www.sapientia.uji.es | 13



Ampliación de Estadística para la Ingeniería Técnica en Informática de Gestión

Irene Epifanio López
Pablo Gregori Huerta

Ampliación de Estadística para la Ingeniería Técnica en Informática de Gestión

Irene Epifanio López
Pablo Gregori Huerta



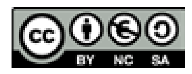
Departament de Matemàtiques

■ Codi assignatura IG23

Edita: Publicacions de la Universitat Jaume I. Servei de Comunicació i Publicacions
Campus del Riu Sec. Edifici Rectorat i Serveis Centrals. 12071 Castelló de la Plana
http://www.tenda.uji.es e-mail: publicacions@uji.es

Col·lecció Sapientia, 13
www.sapientia.uji.es

ISBN: 978-84-692-4538-5



Índice general

Prefacio

Este material va dirigido principalmente a los estudiantes de la asignatura IG23 Ampliación de Estadística de la Ingeniería Técnica en Informática de Gestión de la Universitat Jaume I. Este material se encuadra dentro del programa UJI Virtual, por lo que también se ofrece en abierto a través de Internet, para que pueda ser utilizado por cualquier persona.

Esta asignatura es una materia troncal, que cuenta con una asignatura previa (IG12 Estadística). Dado que constituyen las únicas materias con contenidos estadísticos que cursarán los estudiantes, junto con el hecho de que la titulación sea una ingeniería, hace que el aspecto práctico y aplicado cobre gran valor, pues además el tiempo disponible es escaso en comparación con los conceptos a tratar.

Por estas razones, que también vienen respaldas por profesionales de prestigio reconocido (basta leer los prefacios de libros como Montgomery y Runger [49], Cao *et al.* [11], Vilar [79], Dougherty [24], Moore [51], Devore [22], Mendenhall y Sincich [45], entre otros muchos), el enfoque de la asignatura es el de una estadística aplicada, con que puedan resolver los problemas del mundo real con los que puedan enfrentarse, y no plantear un curso de «estadística matemática elemental», que sería acorde a otro tipo de titulaciones.

En este material se presenta la teoría acompañada de ejemplos extraídos de problemas de exámenes, principalmente, todos ellos relacionados con el campo de la informática, junto con los comandos que usaríamos en el lenguaje R, para obtener los resultados.

Hay un primer capítulo, donde se repasa los puntos fundamentales, vistos en la asignatura previa IG12 (véase [34]), para luego ya centrarnos en los contenidos del programa de esta asignatura, IG23: Inferencia estadística. Estimación; Contrastes de hipótesis; Control de calidad; Diseño de experimentos.

Entendemos que es fundamental trabajar esta materia mediante la realización de tareas, principalmente, la realización de problemas. Las actividades de aprendizaje, sin embargo, no se incluyen en este libro, sino que su lugar está en el aula virtual, tal y como se señala en el programa UJI Virtual. De todas formas, sobre todo de cara al autoaprendizaje, material que también puede complementar parte de la materia de esta asignatura (los dos primeros temas), mediante cuestiones de verdadero/falso, a completar y de elección múltiple es [43].

Es indiscutible la importancia del uso de las clases de ordenador para la enseñanza de la Estadística en la actualidad. Por ello, hay un capítulo dedicado al *software* libre R. De nuevo, no se han incluido las actividades de aprendi-

zaje, que se disponen en el aula virtual. Sin embargo, los estudiantes también pueden disponer de material con actividades en [26].

Se incluye también un formulario, que recopila las fórmulas más importantes tratadas en la asignatura.

En el último capítulo, se realiza igualmente un breve repaso sobre diverso material, que puede ser de interés, para el aprendizaje de la materia.

Por último, nos gustaría agradecer a nuestras familias, la paciencia y el apoyo, que esperemos siga siendo inagotable.

Ahora sólo queda esperar y desear que este material sirva para su fin, ayudar en el aprendizaje y en la resolución de problemas de la Estadística.

Irene Epifanio
Pablo Gregori

Universitat Jaume I

PARTE I
REPASO

Capítulo 1

Repaso previo

El único antídoto para esta posible manipulación y para participar efectivamente en la argumentación pública basada en cifras y datos, consustancial a la vida democrática, es un conocimiento básico de los métodos estadísticos. En este sentido, una formación en los conocimientos estadísticos básicos es necesaria para cualquier ciudadano.

Daniel Peña

Este es un tema de repaso que recapitula las ideas básicas que ya se vieron previamente el curso anterior en IG12 Estadística. Sólo pretende refrescar la memoria, centrar ideas y tener una visión global en unas pocas páginas. En el libro de Gregori y Epifanio [34], podéis encontrar material más detallado que el presente tema, ya que trataba la asignatura previa a ésta.

1.1. Introducción

A continuación, se presentan varios ejemplos del tipo de problemas que seremos capaces de resolver al final del curso.

Ejemplo 1.1: En una empresa, se realiza diariamente un control sobre el número de intentos de acceso fraudulentos a cuentas de los trabajadores de la empresa. El control se realiza a partir de una muestra de 500 intentos de acceso, seleccionados aleatoriamente del total de intentos de acceso diario. Los intentos de acceso se clasifican sencillamente en «buenos» o «malos» según si la contraseña escrita al intentar acceder es correcta o no. En teoría se considera que la tasa de intentos de acceso fraudulentos no ha de superar el 2 % del total de intentos. Supongamos que hoy, de los 500 intentos de acceso de la muestra, 12 han sido fraudulentos, es decir, un 2.4 %. ¿Tenemos motivos suficientes para sospechar que alguien está intentando acceder fraudulentamente al sistema o se debe únicamente al azar?

Ejemplo 1.2: Estamos interesados en comparar los tiempos de ejecución de 5 algoritmos de ordenación (algoritmo de la burbuja, de selección, de inserción, *quicksort*, *treesort*) para un cierto tipo de datos de un tamaño determinado y con un cierto grado de desorden. Para ello, consideramos diversos conjuntos de entrada de entre los que estamos interesados y obtenemos el tiempo de CPU

de ejecución con cada algoritmo. Algunas preguntas que querríamos contestar en base a los resultados obtenidos podrían ser: ¿Existe diferencia significativa entre los 5 algoritmos? ¿Hay un algoritmo mucho mejor que los otros? ¿Pueden clasificarse los algoritmos en diversos grupos homogéneos en el sentido que dentro de cada grupo no difieran significativamente?

El problema podría complicarse si, por ejemplo, el tamaño de los datos a ordenar o el grado de desorden no son siempre los mismos, entonces deberíamos plantear un modelo adecuado al problema.

Ejemplo 1.3: Se pretende diseñar un ratón ergonómico para niños de 7 a 9 años. Hemos de conocer la forma de su mano derecha por lo que hemos de tomar distintos datos antropométricos de un conjunto de niños. Supongamos que estamos interesados en la longitud de su dedo índice. Realizamos un estudio piloto con 30 niños, de los que obtenemos una media de 6 cm y una desviación típica de 0.4 cm. Si deseamos poder afirmar con un 95 % de confianza que la media es imprecisa como mucho en 0.1 cm, ¿cuántos datos deberíamos tomar? Una vez tomados, podríamos calcular un intervalo de confianza al 95 % para la media.

Ejemplo 1.4: Este ejemplo se escapa de los objetivos del curso, pero muestra otro tipo de problemas que pueden resolverse utilizando la Estadística. Desearíamos diseñar un detector automático del tan molesto correo basura (*spam*), de forma que se filtrara este correo antes que colapsara los buzones de los usuarios. Utilizando la información de 5000 e-mails, se intentará predecir si un nuevo correo electrónico es correo basura o no, de manera automática. Por ejemplo, variables que podrían sernos útiles serían el porcentaje de aparición de determinadas palabras, como puede ser: «free», «help», «Irene», etc. Al final se podrían obtener (mediante métodos que no veremos) reglas como:

si (% Irene $<$ 0.6) & (% help $>$ 1.5) entonces *spam*
si no *e-mail*

Veamos ahora de qué se encarga la Estadística. La ciencia Estadística tiene un doble objetivo:

- La generación y recopilación de datos que contengan información relevante sobre un determinado problema (muestreo).
- El análisis de dichos datos con el fin de extraer de ellos dicha información. El primer paso en el análisis de los datos consistirá en describirlos a través de ciertas medidas y gráficas, lo cual nos facilitará su comprensión (estadística descriptiva). Sin embrago, buscamos ir más allá y poder sacar conclusiones basadas en dichos datos. Para ello, podremos recurrir a plantear un modelo matemático (teoría de la probabilidad) que nos permitirá después extraer las conclusiones que nos interesan (inferencia estadística).

Por tanto, un modelo estadístico constará de varias partes: *a*) muestreo (apartado 1.7), *b*) estadística descriptiva (apartado 1.2), *c*) confección de un modelo matemático (teoría probabilidad) (apartados 1.4, 1.5, 1.6), *d*) inferencia estadística (este curso). Esta última parte (d) se considerará en este curso, mientras que las restantes se han tratado en la asignatura IG12 Estadística (o F04 para los procedentes del viejo plan de estudios).

En resumen, la Estadística estudia los métodos científicos para recoger (hacer un muestreo), organizar, resumir y analizar datos (estadística descriptiva), así como para obtener conclusiones válidas (inferencia estadística) y tomar decisiones razonables basadas en tal análisis.

Así, en el ejemplo 1.2, primero tomamos una muestra aleatoria de entre *todos* los archivos de ese tipo (tamaño y grado de desorden), obtenemos los tiempos de ejecución con cada algoritmo, después se describirían (medias, varianzas, gráficos, ...), se propondría un modelo adecuado y obtendríamos las conclusiones de interés (respuestas a las preguntas planteadas).

Repasemos ahora algunos conceptos básicos:

Población: Conjunto de todos los individuos que son objeto de estudio y sobre los que queremos obtener ciertas conclusiones. Ejemplos:

- Todos los niños entre 7 y 9 años (ejemplo 1.3).
- Todos los e-mails recibidos y por recibir (ejemplo 1.4).

Como puede verse, a veces las poblaciones existen físicamente y son finitas aunque muy grandes, en cambio otras veces la población es de carácter abstracto. En general, en lugar de hacer un estudio de todos los elementos que componen la población (hacer un censo), se escoge un conjunto más reducido.

Muestra: Es un subconjunto, una parte de la población que seleccionamos para un estudio.

Es deseable que la muestra extraída «se parezca» a la población, es decir,«que sea como la población pero en tamaño reducido». El objetivo es que la muestra sea representativa de la población. Notemos que si la muestra es mala, las conclusiones extraídas no serán válidas, podrían ser erróneas.

Ejemplo 1.3: Si para obtener medidas para el ejemplo 1.3 acudiéramos a un entrenamiento de baloncesto de niños entre 10 a 11 años, ¿obtendríamos una muestra representativa de la población o sesgada?

Es obvio que estará sesgada.

Tamaño muestral: Es el número de observaciones de la muestra, N.

Variable aleatoria: Es una característica aleatoria que podemos expresar numéricamente, es la característica que estamos midiendo en cada individuo. Una característica aleatoria será una característica que tomará un valor para cada individuo.

Las variables aleatorias las denotaremos con letras mayúsculas: X, Y, ...

Las variables aleatorias pueden clasificarse en:

- Cualitativas o categóricas: expresan una cualidad.
- Cuantitativas: tienen propiamente carácter numérico.

Variables cualitativas: Las variables cualitativas a su vez se subdividen en: ordinales o no ordinales, según si las categorías pueden o no disponerse bajo un orden con sentido.

Ejemplos de variables cualitativas no ordinales:

- Distribución de linux: 1 = Red Hat, 2 = Suse, 3 = Debian, 4 = Otras
- Mail: 1 = SPAM, 0 = No SPAM
- Sexo de una persona: 1 = Mujer, 2 = Hombre
- Adicción al tabaco: 1 = Fuma, 2 = No fuma
- Tipo de defectos de un frigorífico defectuoso: 1 = Termostato, 2 = Compresor, 3 = Motor, 4 = Cableado, 5 = Revestimiento, 6 = Otros
- **Ejemplo 1.5:** Los alumnos de 3° ITIG quieren irse de viaje de fin de curso para celebrar que han aprobado y para sacarse unos euros deciden vender gorras. Quieren conocer el color preferido por los compradores potenciales, por tanto, les interesa la variable aleatoria: «Color de la gorra preferido por los miembros de la UJI», con posibles valores: 1 = Negro, 2 = Blanco, 3 = Rojo, 4 = Otros.

Ejemplos de variables cualitativas ordinales:

- Interés sobre una determinada materia: 1 = Bajo, 2 = Medio, 3 = Alto
- Cualquiera de las de la encuesta de evaluación de la docencia: 1 = Muy desfavorable, 2 = Desfavorable, 3 = Indiferente, 4 = Favorable, 5 = Muy favorable

Las **variables cuantitativas** también se dividen en dos:

- Discretas: toman valores discretos, es decir, en un conjunto numerable (podemos contar los posibles valores que pueden adoptar). Existen «espacios» entre los posibles valores que puede adoptar la variable.
- Continuas: como indica su nombre, toman valores en un conjunto no numerable. Los valores que adoptan estas variables, pueden estar tan cercanos como se quiera.

Ejemplos de variables discretas:

1. Número de piezas defectuosas en un lote de 100 piezas.
2. Número de caras obtenidas al lanzar una moneda 20 veces.
3. Número de cincos al lanzar un dado 60 veces.

En los tres casos anteriores los valores que pueden adoptar son finitos: en 1) de 0 a 100, en 2) de 0 a 20, en 3) de 0 a 60. Sin embargo, podría no ser así, podría adoptar valores discretos no limitados:

1. Número de errores en una superficie de grabación magnética.
2. Número de mensajes que llegan a un servidor en una hora.
3. Número de manchas de más de 1 mm^2 en una lámina.
4. Número de defectos en 2 m de cable.
5. Número de veces al mes que va al cine un estudiante de ITIG.

Ejemplos de variables continuas:

1. **Ejemplo 1.2:** Tiempo de ejecución del algoritmo de la burbuja para el tipo de archivos considerado.
2. **Ejemplo 1.3:** Longitud de la mano de niños de 7 a 9 años.
3. Peso de ciertas piezas.
4. Tiempo de vida (duración) de ciertos motores.
5. Dureza de cierto material.
6. Resistencia de cierto producto.
7. Notas de estudiantes de ITIG.
8. Euros gastados con el móvil en un mes por un estudiante de la UJI.

Observación: La distinción entre variables continuas y discretas no es rígida. Las variables continuas anteriores corresponden a medidas físicas que siempre pueden ser redondeadas, por ejemplo, la longitud podemos medirla hasta el milímetro más cercano o el peso hasta el gramo más cercano. Aunque estrictamente hablando, la escala de dichas medidas sea discreta, las consideraremos continuas como una aproximación a la verdadera escala de medida.

Resumiendo, las variables aleatorias pueden ser:

1. Categóricas o cualitativas
   a) No ordinales
   b) Ordinales

2. Cuantitativas

   a) Discretas

   b) Continuas

1.2. Descripción de una muestra

Para describir una muestra, podemos valernos de tablas de frecuencias, de métodos gráficos (histogramas, diagramas de cajas, etc.) y medidas descriptivas.

Recordémoslas brevemente, ayudándonos del ejemplo siguiente.

Ejemplo 1.6: Tabla de frecuencias de las notas del grupo A de la asignatura IG23 en febrero de 2003.

Límites de clase	(Marca de clase)	Frecuencia absoluta	Frecuencia relativa	Frecuencia acumulada	Frecuencia rel. acumulada
[0, 2.5)		11			
[2.5, 5)		29			
[5, 7.5)		51			
[7.5, 10]		29			

Frecuencia absoluta: Número de observaciones en el intervalo.

Frecuencia relativa: Número de observaciones en el intervalo / tamaño muestral; suma 1; indica el porcentaje de observaciones en el intervalo.

Frecuencia acumulada: Suma de las frecuencias de los intervalos anteriores, incluyendo el actual. Indica el número de observaciones por debajo del extremo superior de la clase. Obviamente, el último valor es el tamaño muestral.

Frecuencia relativa acumulada: Frecuencia acumulada/tamaño muestral. Indica el porcentaje muestral por debajo del extremo superior de la clase. El último valor será 1 (100 %).

Normalmente, las clases son de igual anchura, pero podrían no serlo:

Intervalo	Frec. abs.	Frec. rel.	Frec. acum.	Frec. rel. acum.
[0, 5)	40			
[5, 7)	42			
[7, 9)	20			
[9,10]	18			

Los gráficos nos permiten también ilustrar la distribución de los datos.

Histograma: Pueden ser de frecuencias absolutas, relativas, acumuladas o relativas acumuladas, según que represente la altura de la barra.


Ejemplo 1.6: Ejemplo de histograma para las notas de otro grupo.

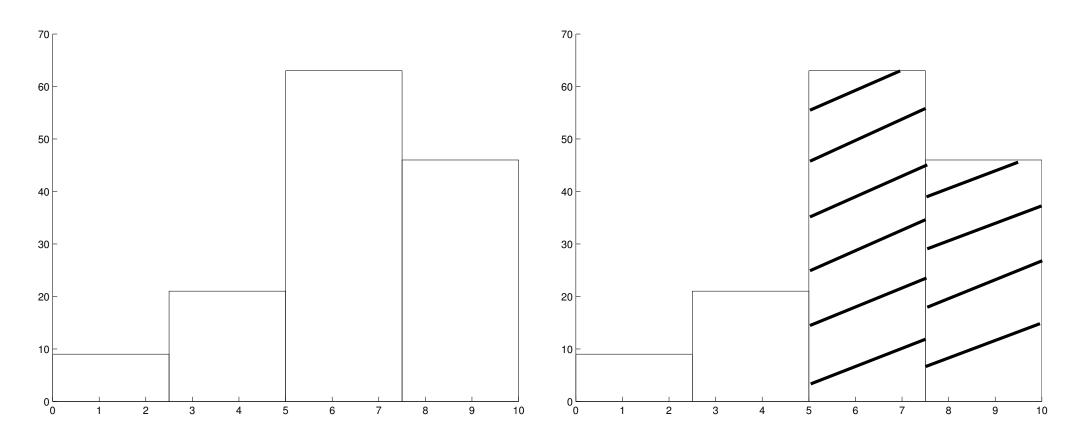


Figura 1.1: Histograma de frecuencias absolutas

Los histogramas nos muestran cómo se distribuyen (cómo se reparten) los datos, las cimas de las barras indican la forma de la distribución. Además, el área de cada barra es proporcional a la correspondiente frecuencia.

Ejemplo 1.6: El área rayada del histograma anterior es el 78.4 % del área total de todas las barras, por tanto, el 78.4 % de las notas están en las correspondientes clases, o sea, el 78.4 % de las notas están entre 5 (inclusive) y 10.

Hay muchos más métodos gráficos: diagramas de barra, de sectores, polígonos de frecuencias, diagrama de cajas (*boxplot*), Pareto, etc.

Además de las gráficas, otra forma de resumir los datos es mediante medidas descriptivas numéricas, que podemos dividir en:

- Medidas de posición o centrales: dan cuenta de la posición de las observaciones
- Medidas de dispersión: indican la dispersión (variabilidad) de los datos
- Medidas de forma: miden la forma de distribuirse los datos

Medidas de posición: media, mediana, moda y percentil.

Media: Si tenemos una muestra $\{x_1, x_2, ..., x_N\}$,

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}. \tag{1.1}$$

Calculadora: $\boxed{\overline{x}}$

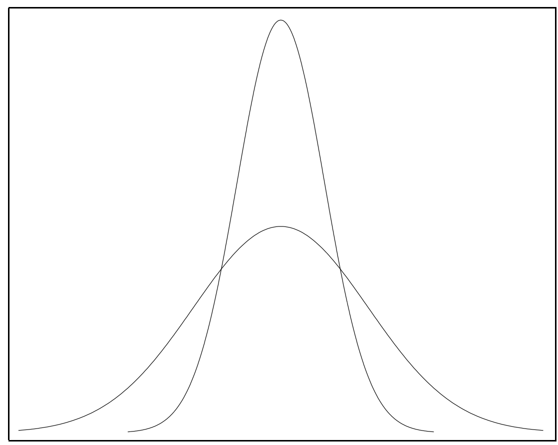
Ejemplo 1.7: Nota media de 5 prácticas: $\{10, 8, 9, 7, 9\}$

Es 43/5= 8.6.

Una medida de posición no es suficiente para describir los datos, porque no informa acerca de la variabilidad de los datos.

Ejemplo 1.8: La nota media de prácticas es 5.2 tanto para $\{0, 2, 5, 9, 10\}$ como para $\{4, 5, 5, 6, 6\}$, sin embargo, claramente su dispersión es distinta.

Si representamos los histogramas mediante curvas continuas, apreciaremos la distinción entre posición y dispersión.



Misma posición y diferente dispersión

Distinta posición y misma dispersión

Medidas de dispersión: rango, rango intercuartílico, varianza, desviación típica o estándar, coeficiente de variación.

Varianza:

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1} = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_N - \overline{x})^2}{N-1} \qquad (1.2)$$

Fórmula alternativa:

$$s^2 = \frac{\sum_{i=1}^{N} {x_i}^2 - N \cdot \overline{x}^2}{N-1} = \frac{{x_1}^2 + {x_2}^2 + \cdots + {x_N}^2 - N \cdot \overline{x}^2}{N-1} \qquad (1.3)$$

Calculadora: $\boxed{\sum x^2}$, $\boxed{\overline{x}}$ o bien $\boxed{\sigma_{N-1}}$, $\boxed{x^2}$

Observación: comprobación de la fórmula alternativa

$$\sum_{i=1}^{N}(x_i - \overline{x})^2 = \sum_{i=1}^{N}(x_i^2 - 2 \cdot \overline{x} \cdot x_i + \overline{x}^2) = \sum_{i=1}^{N} x_i^2 - 2 \cdot \overline{x} \cdot \sum_{i=1}^{N} x_i + N \cdot \overline{x}^2 =$$

$$\sum_{i=1}^{N} x_i^2 - 2 \cdot N \cdot \overline{x}^2 + N \cdot \overline{x}^2 = \sum_{i=1}^{N} {x_i}^2 - N \cdot \overline{x}^2$$

Por la fórmula 1.2 puede apreciarse que a mayor varianza, mayor dispersión, pues calculamos desviaciones de la media al cuadrado. Por esto último (cuadrados), **la varianza siempre será mayor o igual que cero**. Recordad: nunca negativa, siempre positiva.

¿Por qué dividir por N - 1, en lugar de por N? Por razones técnicas que ya se comentarán más adelante; una justificación intuitiva sería considerar el caso en que N=1 (un único valor muestral). Si N es grande no habrá apenas diferencia.

Ejemplo 1.3 (continuación): si sólo observáramos 1 niño (N=1) y nos diera como medida 7 cm, ¿cuál sería s^2? ¿Y si dividiéramos por N?

Si dividimos por $N-1$, no podemos obtener s^2, que es bastante coherente dado que con un único dato difícilmente podemos conocer la variación. □

La varianza es muy apropiada por ciertas propiedades (si dos variables son independientes, la varianza de la suma es la suma de las varianzas), pero tiene un problema: cambia las unidades de los datos, ya que hacemos un cuadrado. Para resolverlo se usa la raíz cuadrada de la varianza:

Desviación típica o estándar:

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}} = \sqrt{s^2} \qquad (1.4)$$

Calculadora: $\boxed{\sigma_{N-1}}$

1.3. Descripción de la población

Hasta ahora hemos examinado diversas formas de describir una muestra. Aunque la descripción de un conjunto de datos es de interés por sí misma, muchas veces lo que se pretende es generalizar y extender los resultados más allá de la limitación de la muestra. La población es realmente el foco de interés.

Como ya vimos, el proceso de sacar conclusiones sobre una población basándonos en las observaciones de una muestra de dicha población, es la inferencia estadística.

Puesto que las observaciones se realizan únicamente en la muestra, las características de la población nunca se conocerán exactamente. Para poder inferir («deducir, concluir, tomar decisiones») de una muestra a la población, necesitaremos un lenguaje (paralelo al muestral) para describir la población.

Variables categóricas: Podemos describir la población simplemente indicando la proporción de la población en cada categoría.

Ejemplo 1.5 (continuación): Supongamos hipotéticamente que podemos preguntar a todos los miembros de la UJI:

Población: todos los miembros de la UJI Color	p	La muestra de alumnos de 3º ITIG Frecuencia relativa ($\hat{p}$)
1 = Negro	0.57	0.52
2 = Blanco	0.14	0.07
3 = Rojo	0.09	0.13
4 = Otros	0.2	0.28

La proporción muestral de una categoría es una *estimación* de la correspondiente proporción poblacional (en general desconocida). Puesto que no tienen porqué ser iguales (aunque sí que querríamos que fuesen cuanto más iguales mejor), las denotaremos con letras diferentes:

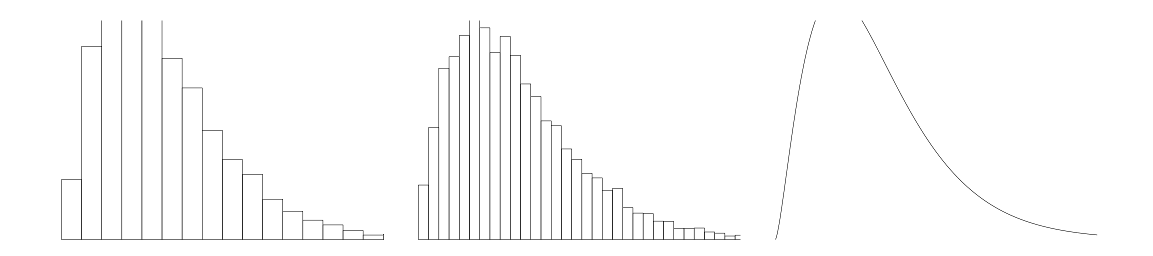
- p = proporción de la población
- $\hat{p}$ = proporción de la muestra

Variables cuantitativas: Para variables cuantitativas, la media, varianza, desviación típica, etc., son descripciones de la población. Estas cantidades se aproximarán con los datos muestrales y constituirán una *estimación* de las correspondientes cantidades para la población. La media de la población la denotaremos mediante la letra μ, la varianza y desviación típica de la población con σ^2 y σ respectivamente. Recordemos que la media muestral era $\overline{x}$, la varianza muestral, s^2 y la desviación típica, s. Notemos que $\overline{x}$ es una *estimación* de μ (desconocida) y s es una estimación de σ (desconocida).

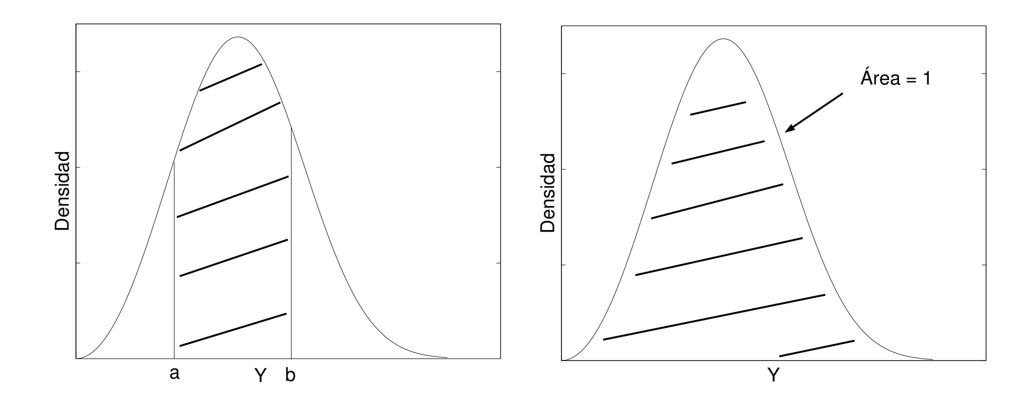
Nota que las cantidades poblacionales las denotamos con letras griegas que se corresponden con las respectivas letras latinas, para las cantidades muestrales.

Ejemplo 1.3 (continuación): Con la muestra de 30 niños obtenemos $\overline{x}$ = 6 y s = 0.4. La media de la población (todos los niños entre 7 y 9 años) la llamamos μ y no la conocemos. La desviación típica de la población (todos los niños entre 7 y 9 años) la llamamos σ y no la conocemos.

El histograma también es una buena herramienta que nos informa sobre la distribución de frecuencias de la población. Si, además, la variable es continua, podemos emplear una curva suave para describirla. Esta curva puede verse como una idealización del histograma con clases muy estrechas. Esta curva que representa la distribución de frecuencias, es la **curva de densidad**.

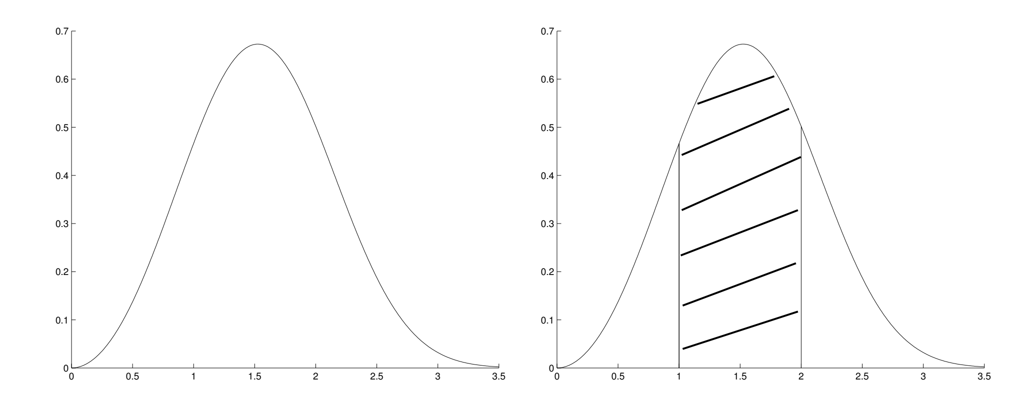


Interpretación de la densidad: El área bajo la curva de densidad entre los valores a y b equivale a la proporción de valores de la variable Y entre a y b.



Debido a la forma en que la curva es interpretada, el área bajo la curva entera debe ser igual a 1.

Ejemplo 1.9: Supongamos que nos interesa la variable X = tiempo (en decenas de miles de horas) de vida de cierta clase de ventiladores de ordenador y que se distribuye según la siguiente curva de densidad:



El área rayada es igual a 0.61, lo cual indica que el 61 % de los valores de la variable están entre 1 y 2.

Para calcular las áreas bajo las curvas de densidad, necesitaríamos integrar, aunque en muchos casos usaremos tablas.

Observación: ¿Cuál sería la frecuencia relativa de un valor concreto, por ejemplo 6 cm, de la variable del ejemplo 1.3? La respuesta es cero (el área es cero). Aunque parezca extraño que la frecuencia relativa de una longitud igual a 6 cm sea cero, pensemos un poco. Si estamos midiendo hasta el milímetro más cercano, entonces, en realidad estamos preguntando la frecuencia relativa entre 5.95 cm y 6.05 cm, que no es cero. Pensemos en la longitud como una variable continua idealizada. Es similar al hecho de que una línea de 1 m, está compuesta de puntos, cada uno de ellos de longitud cero.

En resumen, una medida numérica calculada a partir de los datos es un estadístico. La correspondiente medida numérica que describe la población es un parámetro. En la siguiente tabla se recogen las más importantes:

Medida	Muestral (estadístico)	Poblacional (parámetro)
Proporción	$\hat{p}$	p
Media	$\overline{x}$	μ
Desviación típica	s	σ

1.4. Probabilidad

¿Por qué hemos de estudiar la probabilidad? Las conclusiones de los análisis estadísticos de datos vendrán generalmente dadas en términos probabilísticos (como ya se verá posteriormente en este curso, hasta ahora en el apartado anterior nos hemos limitado a describir los datos). La probabilidad entra en los análisis estadísticos, no únicamente porque el azar influya en los resultados de un experimento, sino también a causa de los modelos teóricos que se usarán en la parte de inferencia estadística. Para poder extraer conclusiones sobre la población, a partir de los datos de una muestra, será necesario recurrir a un modelo matemático (un esquema teórico de comportamiento) que nos determine las reglas de inferencia que es necesario utilizar. La probabilidad es el lenguaje y la fundamentación matemática de la estadística inferencial, de igual manera que las reglas de la gramática proporcionan las bases para organizar ideas a partir de las palabras que forman la lengua.

Espacio muestral y puntos muestrales: El espacio muestral S de una variable aleatoria X es el conjunto de valores que puede tomar dicha variable. Cada uno de los elementos de S se llama punto muestral.
Suceso: es un subconjunto A de S.

Una probabilidad es una cantidad numérica que expresa la verosimilitud de un cierto suceso A (*certidumbre de que el suceso A ocurra*), denotada como $P(A)$ (**probabilidad del suceso A**). Este número estará **siempre** entre 0 y 1 (ambos inclusive). Sólo tiene sentido hablar de probabilidad en el contexto de un **experimento aleatorio**, es decir, una operación (proceso) cuyo resultado viene determinado al menos parcialmente por el azar. De esta forma, cada vez que se lleva a cabo una operación, el suceso A puede ocurrir o no ocurrir. Dicho

de otro modo, un experimento aleatorio es aquel que proporciona diferentes resultados aun cuando se repita siempre de la misma manera.

La probabilidad podemos interpretarla en términos frecuenciales. Así, si un experimento aleatorio se pudiera repetir un número infinito de veces, la probabilidad de un suceso A, $P(A)$, se interpretaría como la frecuencia relativa de la ocurrencia del suceso A en una serie infinita de repeticiones de dicho experimento. O sea, si ese experimento se repitiera un número grande de veces y por cada repetición anotásemos la ocurrencia o no de A, se tendría:

$$\text{P(A)} \longleftrightarrow \frac{\text{número de veces que ocurre A}}{\text{número de veces que se repite el experimento}}$$

donde $\longleftrightarrow$ quiere decir: aproximadamente iguales si el experimento se repite muchas veces.

Ejemplo 1.10: P("sacar cara") = 0.5, podéis lanzar una moneda muchas veces y comprobarla (si no está trucada, ¡claro!). De todas maneras, fíjate que para una realización concreta del experimento, quizá no obtengas exactamente la mitad de las veces cara. De hecho, cada vez que realices el experimento la frecuencia relativa seguramente cambiará, pero tras repetirlo muchísimas veces la frecuencia relativa (empírica o experimental) tenderá hacia la probabilidad teórica del suceso. La aproximación mejorará conforme más repeticiones se lleven a cabo. Las probabilidades de un experimento aleatorio a menudo se asignan sobre la base de un modelo razonable del sistema que se estudia, es decir, se asignarán siguiendo las especificaciones de un modelo teórico que plantearemos (en los dos próximos apartados) y que explicaría el fenómeno que se estudia. Otras veces, nos basaremos en los resultados de estudios realizados.

Recuerda que siempre $0 \leq P(A) \leq 1$, siendo A un suceso.

1.5. Algunos modelos de distribuciones de probabilidad para variables discretas

Recordemos que una **variable aleatoria** es una variable cuyo valor depende del resultado de un experimento aleatorio. En el apartado 1.1 se vieron diversos ejemplos y se distinguió entre variables cualitativas (o categóricas) y cuantitativas. Dentro de éstas últimas, diferenciamos entre:

- **variables discretas**: toman un conjunto finito o infinito numerable (que se pueden contar) de valores.
- **variables continuas**: su espacio muestral está formado por un conjunto infinito de valores que no podemos contar.

A continuación repasaremos algunos modelos matemáticos concretos que nos darán la pauta de variabilidad asociada a una variable aleatoria. Estos modelos matemáticos se llaman distribuciones de probabilidad. Una **distribución de probabilidad** es un conjunto de probabilidades para los posibles

distintos sucesos que pueden darse en un experimento aleatorio, en otras palabras, lo que nos proporciona es cómo se *distribuye* la probabilidad entre los sucesos que pueden producirse.

Nota: el curso pasado sólo visteis el caso univariante (una única variable); sin embargo, también existen modelos que consideran varias variables conjuntamente.

Repasaremos 3 modelos (hay muchos más), que corresponderán a la consideración de experimentos con determinadas características. El fin de estos modelos teóricos es la descripción razonable de algunos fenómenos aleatorios. Son modelos aleatorios o estocásticos, que se diferencian de los modelos matemáticos determinísticos. Para los modelos determinísticos, los resultados se encuentran predeterminados por las condiciones bajo las cuales se verifica el experimento, es decir, dada una entrada, su salida (resultado) queda determinada. Por ejemplo, una fuente de alimentación (E) suministra corriente a un circuito de resistencia eléctrica (R), el modelo matemático que nos describiría el flujo de corriente viene dado por la Ley de Ohm I=E/R. El modelo suministraría el valor de I tan pronto como se dieran los valores de E y R. Sin embargo, para los experimentos aleatorios, los resultados no pueden predecirse con certeza.

Los tres modelos que repasaremos son: la uniforme discreta, la Binomial y la Poisson. Tanto la distribución Binomial como la de Poisson tienen aplicación en fiabilidad y en control de calidad. La fiabilidad estudia la probabilidad de funcionamiento de una unidad, entendida no sólo como parte indescomponible de un sistema, sino también como un sistema o subsistema en sí.

Distribución uniforme discreta: Es la distribución que sigue una variable aleatoria X que toma n posibles valores x_1, x_2, ..., x_n con la misma probabilidad. Por tanto,

$$P(X = x_i) = \frac{1}{n} \qquad \text{i = 1, ..., n}$$

Ejemplo 1.11: X="resultado al lanzar un dado no trucado".

1.5.1. Binomial

Esta distribución tiene una amplia gama de aplicaciones, sobre todo cuando se trata de realizar pruebas cuyo resultado sólo puede adoptar dos valores: «éxito» o «fracaso».

Supongamos que llevamos a cabo un **proceso de Bernoulli**, es decir, una serie de pruebas. Cada prueba puede resultar en un «éxito» o en un «fracaso». La probabilidad de éxito es la misma cantidad, p, para cada prueba, sin importar los resultados de las otras pruebas, o sea, las pruebas son independientes.

La variable aleatoria X que representa el número de éxitos en una serie de n pruebas de un proceso de Bernoulli, tiene una **distribución binomial**.

Ejemplo 1.12: El ejemplo por excelencia de variable aleatoria distribuida como una binomial, sería X = "número de caras obtenidas al lanzar una moneda no trucada 5 (por ejemplo) veces", en este caso n = 5 y p = 0.5. O bien, X = "número de caras obtenidas al lanzar una moneda trucada (de forma que la probabilidad de salir cara sea 0.7) 10 (por ejemplo) veces", en este caso n = 10 y p = 0.7.

Si la variable X sigue (se distribuye como) una distribución binomial de parámetros n y p (siendo n el número de pruebas y p la probabilidad de éxito), que representaremos como $X \sim Bi(n, p)$, las probabilidades se distribuyen de la siguiente manera (considerando combinatoria podría deducirse):

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \qquad x = 0, 1, ..., n, \qquad q = 1 - p, \text{ donde}$$

$$\binom{n}{x} = \frac{n!}{x! \cdot (n-x)!} \qquad \text{siendo} \qquad n! = n \cdot (n-1) \cdot (n-2) \cdot ... \cdot 2 \cdot 1$$

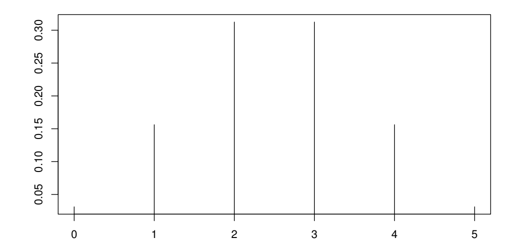
Fíjate que una variable $X \sim Bi(n, p)$, sólo puede tomar un número de valores finito, de 0 a n.

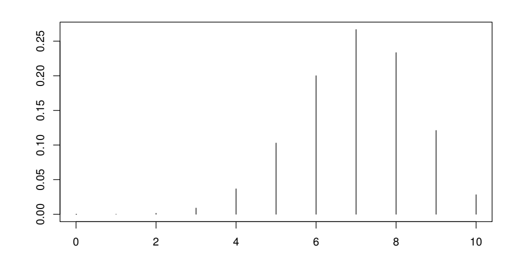
Calculadora: para calcular $\binom{n}{x}$ puede emplearse las teclas $\boxed{nCr}$ o bien $\boxed{x!}$. Recuerda también que $\binom{n}{0} = \binom{n}{n} = 1$, $\binom{n}{1} = \binom{n}{n-1} =$ n, $0! = 1$ y $1! = 1$.

Ejemplo 1.12: Las siguientes gráficas muestran cómo se distribuye (reparte) la probabilidad entre los puntos muestrales, para las dos variables de este ejemplo. Fíjate que si sumamos todas las probabilidades de los puntos muestrales obtendremos 1.



Binomial(10, 0.7)





En cada problema, debe especificarse qué quiere decir «éxito». «Éxito» puede ser «salir cara» como en el ejemplo anterior, o bien, por ejemplo «ser defectuoso», «ser satisfactorio», o «cumplir las especificaciones», etc.

Más ejemplos de variables aleatorias con distribución Binomial son:

- Una máquina-herramienta desgastada produce 1 % de piezas defectuosas. La variable $X =$ "número de piezas defectuosas en las siguientes 50 piezas producidas" seguirá una distribución Binomial, con parámetros $n = 50$ y $p = 0.01$.

- De todos los bits transmitidos a través de un canal de transmisión digital, 10 % se reciben con error. La variable $X =$ "número de bits con error en los siguientes 10 bits transmitidos" se distribuye como una distribución Binomial(10,0.1).

- Un producto electrónico contiene 40 circuitos integrados. La probabilidad de que cualquiera de los circuitos integrados esté defectuoso es 0.02, y los circuitos integrados son independientes. La variable $X =$ "número de circuitos defectuosos de los 40" es Binomial(40,0.02).

Puesto que estamos estableciendo modelos teóricos que describan el comportamiento de ciertas variables aleatorias, también podremos establecer cuál sería la media poblacional, μ, y la varianza poblacional, σ^2, usando estos modelos.

Para una variable Binomial, $X \sim Bi(n, p)$, se tiene $\mu =$ n·p, y $\sigma^2 =$ n·p·q. La media, μ también se llama esperanza matemática.

1.5.2. Poisson

Consideremos ahora una serie de experimentos que consisten en observar el número de ocurrencias de un hecho en un intervalo de tiempo o espacio determinado. Por ejemplo:

Ejemplo: Número de errores en una superficie de grabación magnética.

Ejemplo: Número de mensajes que llegan a un servidor en una hora.

Ejemplo: Número de fallos de un equipo industrial durante 5 años.

Ejemplo: Número de defectos de fabricación por cada 1000 metros de cable.

Una variable aleatoria X sigue una **distribución de Poisson**, si cuenta el número de ocurrencias por unidad de magnitud, cuando:

- el número de ocurrencias en un intervalo de tiempo o del espacio es independiente del número de ocurrencias en otro intervalo disjunto (proceso sin memoria).
- Además, la probabilidad de que haya una sola ocurrencia en un intervalo muy corto es proporcional a la amplitud del intervalo y
- la probabilidad de que haya más de una ocurrencia en un intervalo muy corto es despreciable.

Si la variable X sigue (se distribuye como) una distribución Poisson de parámetro λ ($X \sim \text{Po}(\lambda)$), donde λ indica el número medio de ocurrencias por unidad de magnitud y suele denominarse parámetro de intensidad, las probabilidades se distribuyen de la siguiente manera:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, 3, ... \qquad (x \in \mathbb{N})$$

Fíjate que una variable $X \sim \text{Po}(\lambda)$, puede tomar un número infinito numerable (contable) de valores.

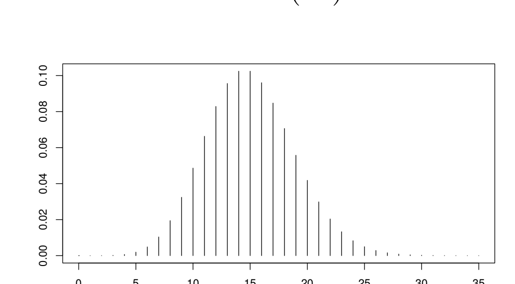
En el caso de una variable Poisson, $X \sim \text{Po}(\lambda)$, se tiene que $\mu = \lambda$ y $\sigma^2 = \lambda$.

Las siguientes gráficas muestran cómo se distribuye (reparte) la probabilidad entre los puntos muestrales, para dos variables Poisson. Fíjate que si sumamos todas las probabilidades de los puntos muestrales obtendremos 1.

Poisson(3)



Poisson(15)



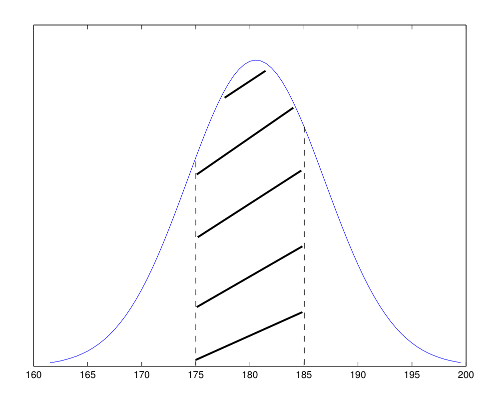
1.6. Algunos modelos de distribuciones de probabilidad para variables continuas

Se recordarán diversos modelos teóricos de distribuciones de probabilidad para variables continuas. En el apartado 1.3, vimos como la distribución de la población de una variable aleatoria **continua** X podría describirse mediante una curva de densidad (como un histograma idealizado), que representaba frecuencias relativas como áreas bajo la curva. Si en un histograma hacemos tender la amplitud del intervalo de clase a cero tendremos un número infinito de intervalos, convirtiéndose el histograma en un número infinito de barras

de grosor infinitesimal, dispuestas de modo continuo (histograma idealizado). De esta forma, llegaríamos a la que llamamos en el apartado 1.3 **curva (o función) de densidad**, y que denotaremos como $f(x)$.

Cada uno de los modelos que repasaremos (y los que no repasaremos) tiene asociado su función de densidad y a través de ella podremos calcular probabilidades de distintos sucesos. La forma de calcular probabilidades para variables continuas difiere de la que se usa para variables discretas. Ahora para calcular la probabilidad de un suceso deberíamos calcular el área comprendida entre el eje x y la función de densidad (o sea, integrar), para los valores señalados por el suceso.

Ejemplo 1.13: Si quisiéramos conocer la probabilidad de que un estudiante de la clase midiera entre 175 y 185 cm, P($175 \leq X \leq 185$), deberíamos calcular el área rayada, es decir, integrar la función de densidad entre 175 y 185 cm.



Según las reglas de probabilidad, tendremos que el **área total bajo la función de densidad es siempre 1**. Además, puesto que la integral de un punto al mismo punto vale cero (el área de una barra con grosor un punto es nula, recuerda también la última observación del punto 1.3), se tiene que para **variables continuas, la probabilidad de que una variable aleatoria continua tome un valor puntual es cero**. Así, en el ejemplo anterior, P($X = 168.96$) = 0, por ejemplo. Por esta razón, para cualquier variable continua X se cumple: P($a \leq X \leq b$) = P($a < X \leq b$) = P($a < X < b$) = P($a \leq X < b$), o sea, para variable continuas (únicamente), la probabilidad será la misma tanto si la desigualdad es o no estricta.

Fíjate que esta última propiedad no se cumple para las variables discretas.

Existe gran cantidad de modelos para variables continuas. Algunos modelos son: la Normal, uniforme, exponencial, Weibull, la t de Student, χ^2 Chicuadrado y F de Snedecor. Cada una de ellas tiene una curva de densidad y viene caracterizada por un/os parámetro/s.

Como ya hemos dicho, para conocer la probabilidad de sucesos para variables continuas deberíamos integrar, sin embargo, para algunos modelos es posible expresar de forma analítica el valor de la integral mediante **la función**

de distribución acumulada que denotaremos $F(x)$ y que nos proporcionará $P(X \leq x)$, es decir, para cada x, la función F nos devolverá la probabilidad de que la variable X tome un valor menor o igual que x. A veces, no existe tal expresión explícita y es preciso recurrir a tablas.

A modo de resumen aclaratorio: cada modelo continuo viene determinado por su función de densidad, f. Hay que tener claro que la función de densidad, f, NO da probabilidades, sino el área bajo dicha función. Para calcular probabilidades hay que usar F, la función de distribución acumulada.

1.6.1. Distribución Uniforme(a,b)

Es la distribución que sigue una variable aleatoria X que toma valores en un intervalo [a,b] con la misma probabilidad. Por ejemplo, las calculadoras científicas con la tecla $\boxed{RAN\sharp}$ o $\boxed{Rnd}$ generan valores aleatorios de una variable uniforme entre 0 y 1. Su función de densidad y su función de distribución tienen la siguiente forma:

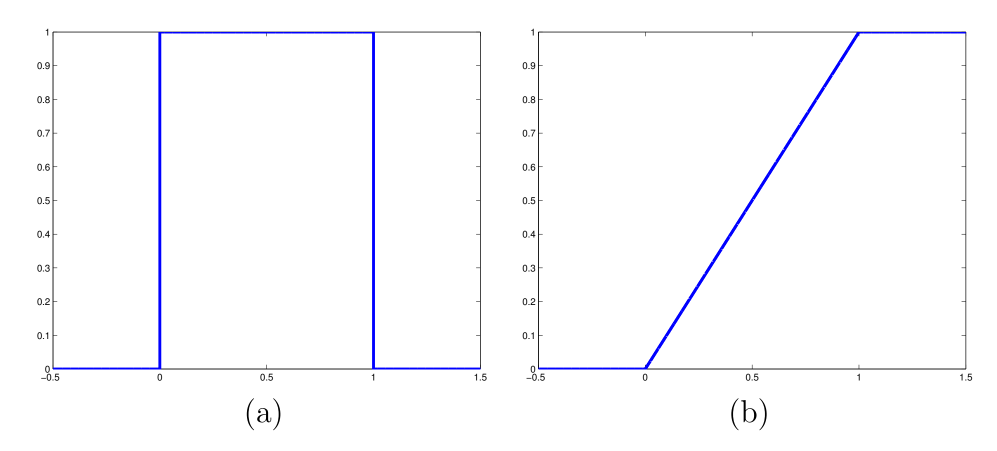


Figura 1.2: (a) Función de densidad, f, de la Uniforme(0,1); (b) Función de distribución, F, de la Uniforme(0,1)

La función de densidad, de distribución acumulada, la media y varianza vienen dadas para una variable Uniforme(a,b) por:

$$f(x; a, b) = \begin{cases} \frac{1}{b-a} & \text{si } a < x < b \\ 0 & \text{en otro caso} \end{cases} \; ; \; \mu = \frac{a+b}{2}, \; \sigma^2 = \frac{(b-a)^2}{12}$$

$$F(x; a, b) = \begin{cases} 0 & \text{si } x < a \\ \frac{x-a}{b-a} & \text{si } a < x < b \\ 1 & \text{si } x > b \end{cases}$$

1.6.2. Distribución Exponencial(λ)

Es usada muchas veces para modelizar el comportamiento de variables aleatorias del tipo "tiempo transcurrido hasta el fallo de un componente industrial" o "tiempo que tarda en completarse un proceso determinado". La función de

densidad y función de distribución de una exponencial de parámetro λ tienen la siguiente forma:

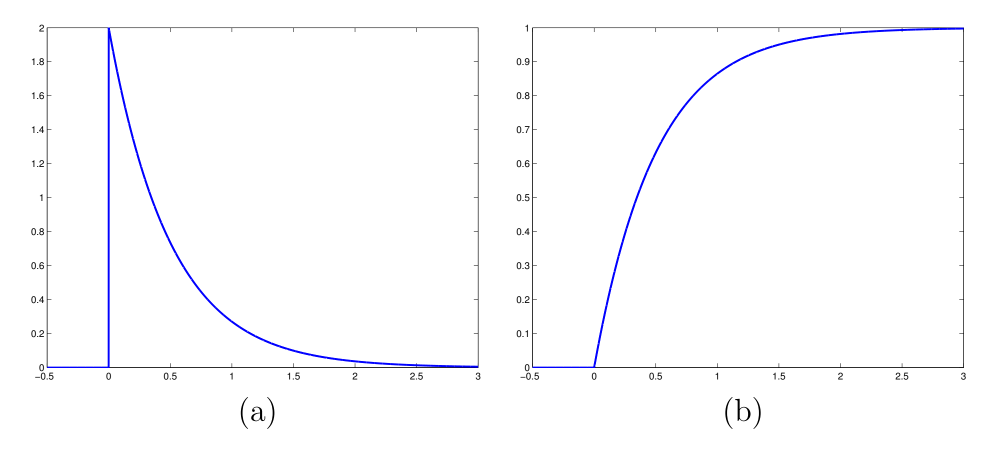


Figura 1.3: (a) Función de densidad, f, de la Exponencial(0.5); (b) Función de distribución, F, de la Exponencial(0.5)

La función de densidad, de distribución acumulada, la media y varianza vienen dadas para una variable Exponencial(λ) por:

$$f(x;\lambda) = \begin{cases} 0 & \text{si } x \leq 0 \\ \lambda e^{-\lambda x} & \text{si } x > 0 \end{cases} \; ; \; \mu = \frac{1}{\lambda}, \; \sigma^2 = \frac{1}{\lambda^2}$$

$$F(x;\lambda) = \begin{cases} 0 & \text{si } x \leq 0 \\ 1 - e^{-\lambda x} & \text{si } x > 0 \end{cases}$$

La distribución exponencial está relacionada con la Poisson de la siguiente forma: si el número de ocurrencias de un determinado fenómeno es una variable con distribución Poisson, el tiempo que pasa entre dos ocurrencias sucesivas es una variable con distribución exponencial.

La distribución Exponencial carece de memoria, se cumple $P(X > s+t|X > s) = P(X > t)$, en el contexto de "tiempos de vida" esto quiere decir que la probabilidad de fallar es independiente del pasado, el sistema no envejece. Aunque pueda parecer algo irreal, no es descabellado por ejemplo suponer que un fusible es "tan bueno como nuevo" mientras esté funcionado.

Más ejemplos de variables aleatorias exponenciales son:

- En una red de computadoras de una gran corporación, el acceso de usuarios al sistema puede modelarse como un proceso de Poisson con una media de 30 accesos por hora. La variable X = "tiempo en horas desde el principio del intervalo hasta el primer acceso" tiene una distribución exponencial con $\lambda = 30$.

- El tiempo entre la entrada de correos electrónicos en una computadora podría modelizarse mediante una distribución exponencial.

- La CPU de un PC tiene un periodo de vida con una distribución exponencial con una vida media de 6.5 años.

1.6.3. Distribución Weibull(α,β)

Otra de las distribuciones que se aplica además de la Exponencial a problemas de fiabilidad y "tiempos de vida de componentes - equipos", es la Weibull(α,β). De hecho, para $\beta = 1$, la Weibull se reduce a la Exponencial. Esta distribución no se vio el curso pasado.

La función de densidad para Weibull(1,β) y distintos valores de β puede verse en el siguiente gráfico, $\beta > 0$ es un parámetro de forma y $\alpha > 0$ de escala.

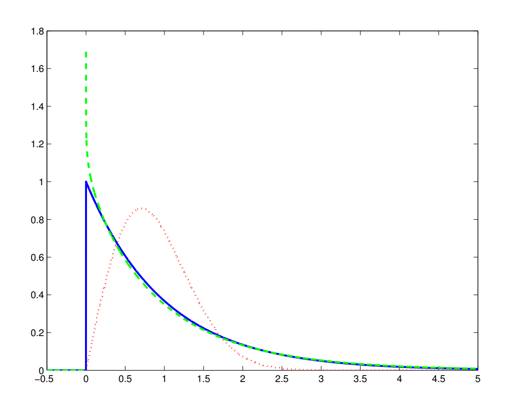


Figura 1.4: En azul y continua: Weibull(1,1), en rojo y puntos: Weibull(1,2), en verde y rayas: Weibull(1,0.95)

A continuación, aparece la expresión de su función de densidad:

$$f(x; \alpha, \beta) = \begin{cases} 0 & \text{si } x \leq 0 \\ \alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}} & \text{si } x > 0 \end{cases}$$

Como ya se ha dicho, la distribución Weibull puede emplearse para modelar el tiempo hasta presentarse un fallo en muchos sistemas físicos diferentes. Los parámetros de esta distribución permiten gran flexibilidad para modelizar sistemas en los que el número de fallos aumenta con el tiempo (por ejemplo, el desgaste), disminuye con el tiempo (algunos semiconductores) o permanece constante (fallos provocados por causas externas al sistema). En la siguiente página http://www.itl.nist.gov/div898/handbook/apr/apr.htm podréis encontrar un capítulo dedicado a la fiabilidad.

Más ejemplos de variables aleatorias Weibull son:

- Tiempo de vida (hasta el fallo) de un chip de memoria.
- Duración de cierto tipo de tubos al vacío.

1.6.4. Distribución Normal(μ,σ^2)

La distribución Normal o Gaussiana es muy importante puesto que se utiliza para modelar muchísimos fenómenos aleatorios; además, incluso se usa para aproximar otras distribuciones. La distribución Normal aproxima lo observado en muchos procesos de medición sin errores sistemáticos, por ejemplo medidas físicas del cuerpo humano (X = "altura de los jóvenes españoles" del ejemplo 1.13, X = "longitud del dedo índice de los niños" del ejemplo 1.3), medidas de

calidad en muchos procesos industriales (como se verá en el tema dedicado al control de calidad), etc. Más ejemplos serían:

- En la detección de una señal digital, el ruido de fondo podría seguir una distribución normal (denominado ruido Gaussiano) con media 0 volts y desviación típica de 0.45 volts.
- El diámetro de los puntos producidos por una impresora matricial seguiría una distribución normal con un diámetro promedio de 0.002 pulgadas y desviación típica de 0.0004 pulgadas.
- La vida de servicio efectiva de baterías usadas en un portátil.
- El volumen de llenado de una máquina automatizada usada para llenar latas de bebida carbonatada.
- La resistencia a la tensión del papel.
- La vida de un componente electrónico bajo condiciones de alta temperatura para acelerar el mecanismo de fallo.
- Voltaje de ruptura de un diodo de un tipo particular.
- Distribución de resistencia de resistores eléctricos, con media 40 ohmios y desviación típica de 2 ohmios.

Una justificación de la frecuente aparición de la distribución Normal es el teorema central del límite: cuando los resultados de un experimento son debidos a un conjunto muy grande de causas independientes que actúan sumando sus efectos, cada uno de ellos de poca importancia respecto al conjunto, es esperable que los resultados sigan una distribución Normal.

Ejemplo 1.3 (continuación): Este ejemplo del ratón ergonómico nos va a permitir ver varios ejemplos más, de variables que podrían suponerse Normales. Para comprobar científicamente las ventajas del ratón ergonómico frente al tradicional, se han realizado diversos estudios. En esos estudios comparativos algunas de las variables empleadas y que podemos suponer Normales son: tiempo de movimiento de cada ratón, actividad eléctrica de varios músculos del antebrazo durante la utilización de cada ratón, intensidad del dolor medida en una cierta escala (VAS).

La función de densidad de una Normal de parámetros μ (media de la población) y σ^2 (varianza de la población, siempre positiva), que denotaremos N(μ,σ^2) (a veces, también se denota N(μ,σ)), tiene la forma siguiente:

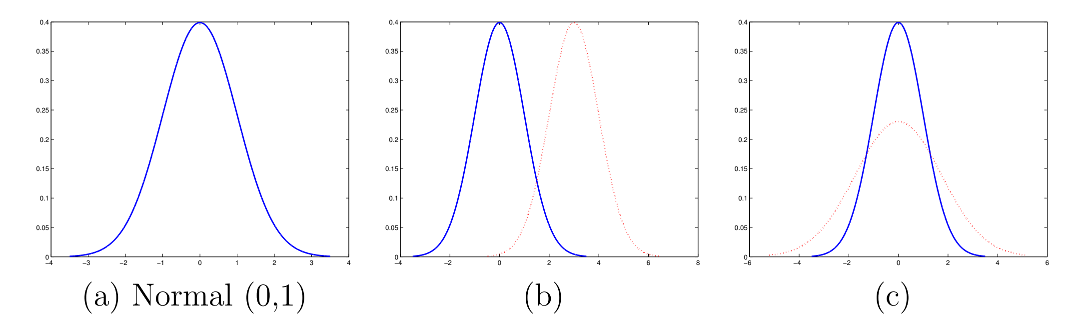

(a) Normal (0,1) (b) (c)

Figura 1.5: (a) Normal(0,1); (b) Un cambio en la media, supone una traslación: Normal(0,1) en azul y continua, Normal(3,1) en rojo y punteada; (c) Un cambio en la varianza, supone un cambio en la variabilidad, pero el área bajo la curva sigue siendo 1, por ello tienen distinta altura: Normal(0,1) en azul y continua y Normal(0,3) en rojo y punteada

Como puede apreciarse, la Normal (campana de Gauss) es simétrica respecto de la media (que en este caso coincide con la mediana y la moda), o sea, el coeficiente de asimetría valdrá cero y además el coeficiente de curtosis es 3.

La función de densidad es:

$$f(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\;e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad , \quad x\in\mathbb{R}$$

La función de distribución acumulada es:

$$F(x;\mu,\sigma^2)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi\sigma^2}}\;e^{-\frac{(y-\mu)^2}{2\sigma^2}}dy$$

La dejamos de esta forma, ya que un integrando de la forma e^{-z^2} no tiene primitiva. Por tanto, para calcularla o bien se emplea algún método numérico o se usan tablas, que es lo que haremos nosotros. Para ello necesitamos presentar la:

Distribución normal estándar: Es aquella distribución normal con media 0 y varianza 1. La denotaremos mediante la letra Z.

Los valores que se recogen en las tablas (las tablas están en el libro de Gregori y Epifanio [34]) son para $N(0,1)$, además algunas calculadoras también permiten calcular probabilidades de una Normal estándar. La tabla nos proporciona:

$$\Phi(z)=P(Z\leq z)=\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}\;e^{-\frac{x^2}{2}}dx \quad , \quad Z\sim N(0,1)$$

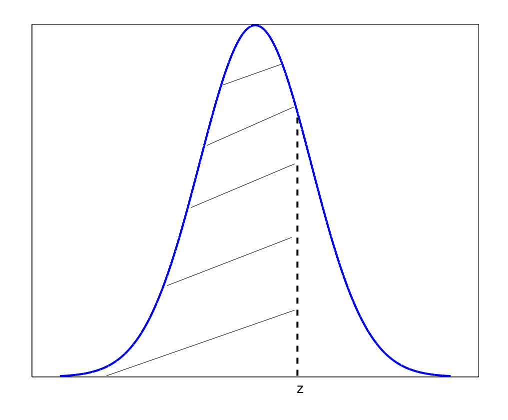


Figura 1.6: $\Phi(z)$

Fíjate que, $P(Z \geq z) = 1$ - $P(Z \leq z)$, $P(Z \leq -z) = 1$ - $P(Z \leq z)$, $P(Z \geq -z) = P(Z \leq z)$. Ayúdate de un gráfico si lo necesitas.

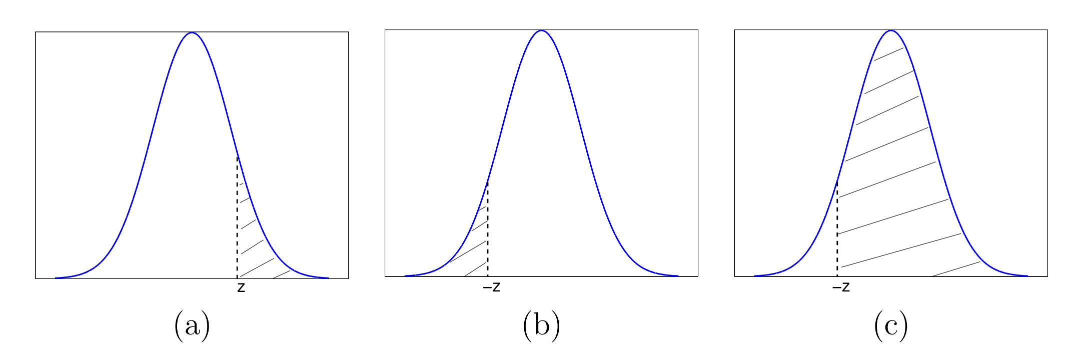


Figura 1.7: (a) $P(Z \geq z)$; (b) $P(Z \leq -z)$; (c) $P(Z \geq -z)$

Con la tabla de la Normal(0,1) podemos calcular cualquier probabilidad de cualquier Normal, con cualquier media μ y varianza σ^2, no necesariamente N(0,1):

Estandarización: sea $X \sim$ N(μ,σ^2), podemos estandarizarla (o tipificarla) y convertirla en una N(0,1) de la siguiente forma:

$$Z = \frac{X - \mu}{\sigma}.$$

O sea, si $X \sim$ N(μ,σ^2), $P(a < X < b) = P(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}) = P(Z < \frac{b-\mu}{\sigma})$ - $P(Z < \frac{a-\mu}{\sigma})$

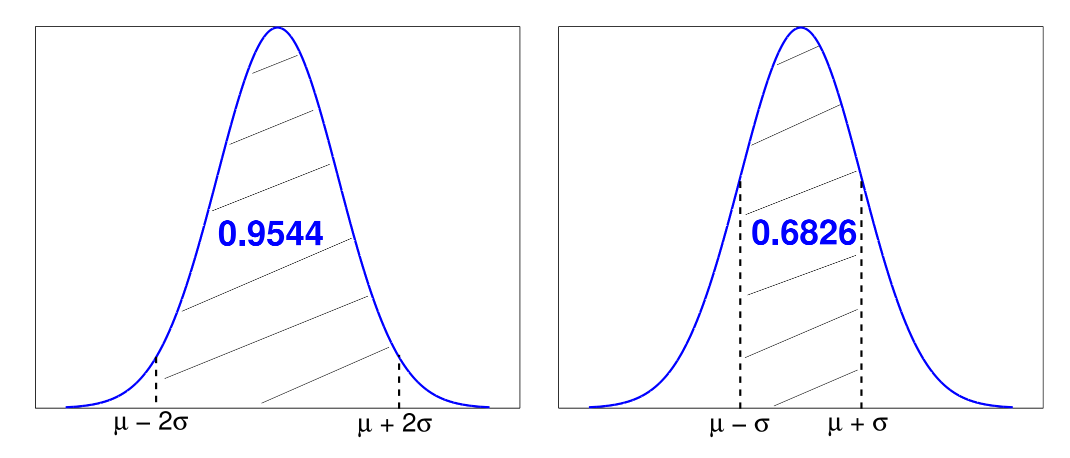
Fíjate que para estandarizar, dividimos por la desviación típica σ, NO por la varianza σ^2.

Fíjate que como la Normal es simétrica respecto su media μ, para $X \sim$ N(μ,σ^2): $P(X \leq \mu) = P(X \geq \mu) = 0.5$. Además, si $x \geq \mu$, $P(X \leq x) \geq 0.5$. También, si $x \leq \mu$, $P(X \leq x) \leq 0.5$. Siempre que tengas dudas, recurre a hacer una representación gráfica.

Ejemplo 1.14: Si $X \sim$ N(μ,σ^2), la fracción (proporción) de números que están a 3 desviaciones de la media es 0.9972, no importa el valor de μ, ni σ^2:

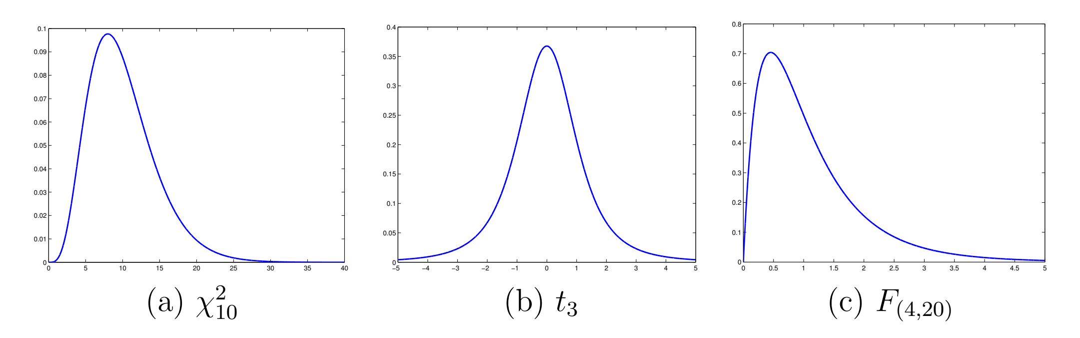
$P(\mu - 3\sigma < X < \mu + 3\sigma) = P(\frac{\mu-3\sigma-\mu}{\sigma} < Z < \frac{\mu+3\sigma-\mu}{\sigma}) = P(-3 < Z < 3)$ $= P(Z < 3)$ - $P(Z < -3) = 0.9986$ - (1 - $P(Z < 3)) = 0.9986$ - (1 - 0.9986) $=$ 0.9972

Puedes comprobar que la fracción de números que están a 2 desviaciones de la media es 0.9544 y la fracción de números que están a 1 desviación de la media es 0.6826.



Observación: Aunque teóricamente la curva normal representa una distribución continua, a veces se usa para aproximadamente describir la distribución de una variable discreta. En esos casos, podría aplicarse una corrección de continuidad, para así obtener una mayor precisión.

Otras distribuciones son la χ^2 Chi-cuadrado, t de Student y F de Snedecor, que usaremos este curso. Un ejemplo de ellas se muestra seguidamente.



(a) χ^2_{10} (b) t_3 (c) $F_{(4,20)}$

Figura 1.8: (a)χ^2 Chi-cuadrado; (b) t de Student; (c) F de Snedecor

1.7. Muestras aleatorias. Otros tipos de muestreo

Recordemos que nuestro objetivo es inferir sobre la población. La población difícilmente puede estudiarse al completo (podría ser económicamente

inviable, temporalmente impracticable, existir sólo conceptualmente, podría también conllevar la destrucción del objeto de estudio, como sería el caso de estudiar el tiempo de vida de una partida de bombillas, etc.). Por ello, nosotros sólo contamos con una muestra de la población. ¿Cómo generalizar más allá de un conjunto de datos particular? El primer paso para el desarrollo de una base para la inferencia estadística es encontrar un modelo probabilístico de las muestras que nos permita utilizarlas para inferir información sobre la población de la que se han extraído: el muestreo aleatorio simple.

Existen diversas técnicas de extracción de muestras de una población (como veremos seguidamente). Nosotros nos centraremos en la más simple:

Muestreo aleatorio simple: Se caracteriza por:

i) cada miembro de la población tiene la misma probabilidad de ser seleccionado;

ii) las selecciones son independientes las unas de las otras.

Ejemplo 1.15: Imaginemos que deseamos conocer el gasto en ocio (en un mes) de los jóvenes (18-30 años) españoles. Para ello extraemos una muestra de tamaño N (por ejemplo $N = 100$) por muestreo aleatorio simple (*pregunto el gasto a N jóvenes completamente al azar*). Si cada estudiante de la clase repitiera el experimento, tendríamos tantas muestras de tamaño N como estudiantes en la clase.

Por tanto, podemos considerar las variables aleatorias X_1, X_2, ..., X_N donde X_1 representa el valor (gasto) de la primera persona elegida (que variará de una muestra a otra), X_2 el valor de la segunda persona, ..., X_N el valor de la N-ésima persona.

Por la condición *i*), la distribución de cada X_i, $1 \leq i \leq N$, es la misma que la de la población (todas las variables X_i siguen la misma distribución). Por *ii*) X_1, X_2, ..., X_N son independientes (el conocimiento de una variable no aporta información acerca de los valores de la otra variable).

En consecuencia, X_1, X_2, ..., X_N, son independientes e idénticamente distribuidas (i.i.d) y constituyen una muestra aleatoria de tamaño N. La función de densidad o probabilidad conjunta de la muestra será por definición: $f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^n f_\theta(x_i)$, donde $f_\theta(x)$ es la distribución de la población y θ el vector de parámetros desconocidos de la misma. Cuando las realizaciones se conocen, $f_\theta(x_1, \ldots, x_n) = L(\theta)$ es una función de los parámetros desconocidos y se denomina *función de verosimilitud*. Esta función será muy útil para hacer inferencias sobre los parámetros.

Estadístico: Es cualquier función de las variables X_1, X_2, ..., X_N que constituyen una muestra aleatoria. Algunos ejemplos son:

Media de muestreo:

$$\overline{X} = \frac{X_1 + X_2 + ... + X_N}{N}$$

Varianza de muestreo:

$$S^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N-1}$$

Un estadístico es una variable aleatoria por ser función de variables aleatorias, por lo cual tiene una distribución que se llama **distribución de muestreo**.

Nota: denotamos con mayúsculas los estadísticos de muestreo por ser variables aleatorias, de esta forma se distinguen de las cantidades muestrales ($\overline{x}$ y s^2, por ejemplo) que vimos en el apartado 1.2, que corresponden a una muestra concreta y tienen un valor numérico concreto.

1.7.1. Distribuciones en el muestreo y estimadores

Se vieron el curso pasado algunas distribuciones en el muestreo: distribución en el muestreo de una proporción, de la media y de la varianza.

- Distribución en el muestreo de la media:

  Sea X_1, X_2, ..., X_N m.a.s. (muestra aleatoria simple) de v.a. (variable aleatoria) X con $E(X) = \mu$ y $Var(X) = \sigma^2$. Media muestral:

  $$\bar{X} = \frac{X_1 + X_2 + ... + X_N}{N}$$

  $E(\bar{X}) = \frac{1}{N}\sum_{i=1}^{N} E(X_i) = \frac{1}{N}\sum_{i=1}^{N} \mu = \mu$

  $Var(\bar{X}) = \frac{1}{N^2}\sum_{i=1}^{N} Var(X_i) = \frac{1}{N^2}\sum_{i=1}^{N} \sigma^2 = \sigma^2/N$

  La distribución exacta de $\bar{X}$ depende de la distribución de la población:

  - Si $X \sim N(\mu, \sigma^2)$, $\bar{X} \sim N(\mu, \sigma^2/N)$
  - Si N grande, $\bar{X}$ puede aproximarse por $N(\mu, \sigma^2/N)$

- Distribución en el muestreo de una proporción. Denotamos por p la proporción desconocida de elementos con cierto atributo y $\hat{P}$ la proporción de elementos de la muestra con dicho atributo:

  $$\hat{P} = \frac{X_1 + X_2 + ... + X_N}{N}$$

  $$X_i \sim Bernoulli(p), E(X_i) = p, Var(X_i) = p(1-p)$$

Es un caso particular de la distribución muestral de una media:

$$E(\hat{P}) = p \quad Var(\hat{P}) = p(1-p)/N$$

La distribución en el muestreo para N grande: $N(p, p(1-p)/N)$

- Distribución muestral de la varianza:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2; \ \sum_{i=1}^{N}(X_i - \bar{X})^2 = \sum_{i=1}^{N}(X_i - \mu)^2 - N(\bar{X} - \mu)^2$$

$E(S^2) = \frac{1}{N-1}(\sum_{i=1}^{N} E[(X_i - \mu)^2] - NE[(\bar{X} - \mu)^2]) = \frac{1}{N-1}(\sum_{i=1}^{N} Var(X_i) - NVar(\bar{X})) = \frac{1}{N-1}(N\sigma^2 - \sigma^2) = \sigma^2$

Si definimos S^2 dividiendo por N: $E(\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2) = \frac{N-1}{N}\sigma^2$

La distribución es, en general, asimétrica, dependiendo de N y la población base. Asintóticamente normal, aunque con aproximación muy lenta.

Para poblaciones normales:

$$\sum_{i=1}^{N}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{N}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{N}}\right)^2$$

  - $\frac{(N-1)S^2}{\sigma^2} \sim \chi^2_{N-1} \Rightarrow E(S^2) = \sigma^2$ y $Var(S^2) = \frac{2}{N-1}\sigma^4$
  - $\bar{X}$ y S^2 son independientes

Especificamos las propiedades deseables de los estimadores:

- *Insesgadez*: Un estimador es insesgado si su distribución muestral está centrada en el parámetro a estimar.
- *Eficiencia*: Dados dos estimadores T_1 y T_2 de un mismo parámetro, se dice que T_1 es más eficiente que T_2 si la varianza de T_1 es menor que la de T_2.
- *Consistencia*: Un estimador es consistente si se aproxima al crecer el tamaño muestral, al valor del parámetro que estima.

Los estimadores que usamos para estimar la media de una población, la varianza y una proporción eran:

- Media: $\bar{X}$
  - Estimador insesgado de μ
  - Estimador consistente de μ
- Proporción: caso particular de $\bar{X}$

- Varianza: S^2
  - Estimador insesgado de σ^2
  - Estimador consistente de σ^2

  Si dividimos por N:

  - Estimador sesgado de σ^2 (asintóticamente insesgado)
  - Estimador consistente de σ^2

Se estudiaron cómo obtener estimadores que, de una manera general, tengan buenas propiedades. En particular se consideraron los siguientes métodos:

- *Método de los momentos.* Este es uno de los métodos más elementales de estimación. En una amplia variedad de problemas, el parámetro desconocido es una función conocida de un número finito de momentos de la distribución. Su estimación por el método de los momentos, consiste en sustituir los momentos de la distribución por los correspondientes momentos muestrales.

- *Máxima verosimilitud.* La idea de la estimación máximo verosímil de los parámetros θ_1, θ_2, ..., θ_k que caracterizan una variable aleatoria X, es elegir los valores de los parámetros que hacen que la muestra observada, x_1, x_2, ..., x_n, sea la más verosímil.

Otros métodos de estimación como el método herramental (*jackknife*) y la estimación autosuficiente (*bootstrap*) puede encontrarse en los apéndices de [56].

Hasta ahora nos hemos limitado a estimar puntualmente un parámetro, no obstante, una estimación puntual podría no ser suficiente, pues no indica el error que cometemos con la estimación. Por esto, sería conveniente dar cierta idea de la precisión de la estimación. Una medida de la precisión que podría usarse sería el *error estándar* del estimador, es decir, la desviación estándar del estimador. Otro enfoque, sería usar un *intervalo de confianza*, donde se espera que esté el parámetro. Este curso, trataremos la estimación por intervalos de confianza.

1.7.2. Otros tipos de muestreo

Aunque a lo largo de este curso siempre supondremos que nuestra muestra se ha obtenido por muestreo aleatorio simple, existen otros tipos de muestreo.

Un objetivo primordial de los procedimientos de muestreo es conseguir que la muestra sea representativa de la población («como la población, pero en tamaño reducido»). Acabamos de presentar el muestreo aleatorio simple, que se usará cuando los elementos de la población sean homogéneos respecto a la característica a estudiar. Pero si disponemos de algún tipo de información sobre

la población sería conveniente emplearla a la hora de seleccionar la muestra. Un ejemplo clásico son las encuesta de opinión, donde los elementos (personas) de la población son (o pueden serlo) heterogéneas en razón a su sexo, edad, profesión, etc. En estos casos interesaría que la muestra tuviera una composición análoga a la población, lo cual se conseguiría mediante muestreo estratificado.

Muestreo estratificado: Los elementos de la población se dividen en clases o estratos. La muestra se toma asignando un número de miembros a cada estrato (pueden usarse distintos criterios: proporcional al tamaño relativo del estrato en la población, proporcional a la variabilidad del estrato, considerando costes, ...) y escogiendo los elementos por muestreo aleatorio simple dentro de cada estrato.

Ejemplo 1.15 (continuación): En este ejemplo, estaría bien dividir los elementos según su nivel económico, y por ejemplo dividirlos según la zona de la ciudad en que habiten: zona centro (clase alta), zona intermedia (clase media), barrios periféricos (clase baja).

Ejemplo 1.16: Queremos conocer la resistencia de los plásticos que hay en un almacén. Los plásticos provienen de dos fabricantes distintos. Sería mejor considerar dos estratos (cada fabricante), que los plásticos como un todo y muestrear sin distinción, porque puede que la distribución sea diferente en cada estrato.

Muestreo por conglomerados: Se utiliza si la población se encuentra de manera natural agrupada en conglomerados, que podemos considerar como una muestra representativa de la población. La muestra se toma seleccionando algunos conglomerados al azar y dentro de ellos analizando todos sus elementos o una muestra aleatoria simple.

Ejemplo 1.15 (continuación): Siguiendo con este ejemplo, dentro de cada estrato (zona de la ciudad) podemos hacer divisiones en calles, las calles serían conglomerados ya que podemos considerarlas homogéneas respecto a la característica a estudiar.

Ejemplo 1.17: Supongamos que queremos analizar el diámetro de unas tuercas que tenemos almacenadas en cajas. Sería más conveniente seleccionar al azar unas cajas y dentro de ellas realizar un muestreo aleatorio simple que llevar a cabo un muestreo aleatorio simple, pues esto implicaría seguramente abrir muchas más cajas.

Las ideas de estratificación y conglomerado son opuestas: la estratificación funciona tanto mejor cuanto mayor sean las diferencias entre los estratos y más homogéneos sean éstos internamente; los conglomerados funcionan si hay muy pocas diferencias entre ellos y son muy heterogéneos internamente.

Muestreo sistemático: cuando los elementos de la población están orde-

nados en listas, se usa el muestreo sistemático. Si la población es de tamaño N y la muestra deseamos que sea de tamaño n, tomaremos k como el entero más próximo a N/n, elegiremos un elemento al azar entre los k primeros, por ejemplo el n_1, después tomaremos los elementos $n_1 + k$, $n_1 + 2k$, etc, hasta completar la muestra.

Como se ha visto en el ejemplo 1.15, los distintos tipos de muestreo pueden emplearse conjuntamente. Por ejemplo, en el análisis de diámetros de tuercas en cajas provenientes de dos fabricantes distintos (juntamos las ideas de los ejemplos 1.16 y 1.17).

Debemos tener *muy* presente que, tan importante es analizar bien los datos suministrados por la muestra, como obtener ésta de forma adecuada. De hecho, un mal diseño muestral, puede llevarnos a conclusiones falsas. Algunos ejemplos de malos diseños muestrales (desgraciadamente de moda), y que por tanto, carecen de validez estadística sería: escoger una muestra de voluntarios (son personas que se "autoseleccionan", en respuesta a un llamamiento general, un ejemplo muy repetido sería el solicitar la opinión sobre un tema en un programa de TV y considerar como muestra las respuestas dadas por teléfono o sms) o el muestreo de conveniencia, donde sólo se seleccionan a los individuos u objetos de más fácil acceso.

Además, debemos tener en cuenta algunas precauciones, para no sufrir algún tipo de sesgo. Éste puede venir de la falta de cobertura (cuando algunos grupos de población se dejan fuera del proceso de selección de la muestra), de la no-respuesta (cuando un individuo seleccionado en la muestra no puede ser localizado o no quiere contestar), del sesgo de respuesta (los encuestados pueden mentir o el encuestador puede también influir en las respuestas), de cómo se hayan redactado las preguntas de la encuesta (puede ser muy influyente), etc.

Por otro lado, las encuestas muestrales son estudios observacionales, es decir, se observa a unos individuos y se mide las variables sin intervenir (influir) en los individuos. En cambio, si lo que se pretende es tratar de establecer alguna relación de causalidad, debería realizarse un experimento (como se verá en el tema dedicado a Diseño de Experimentos).

Observación: El fin de esta aclaración es tratar de dar una visión general y localizar en qué punto del temario nos encontramos, para no perder de vista el objetivo final, que trataremos en este curso. En el ejemplo 1.3 (el del ratón ergonómico para niños), nos interesaba estudiar toda la población de niños. Como eso es inviable, extraeremos una muestra (representativa) de la población, por ejemplo, $N = 100$ niños (muestreo aleatorio simple, apartado 1.7). A partir de esa muestra estudiaremos la variable $X =$ "longitud del dedo índice" en la que estábamos interesados. Esta variable es cuantitativa y continua. (Podía habernos interesado más variables continuas como $Y =$ "longitud entre dos puntos determinados de la mano", u otro tipo de variables, como Z = "satisfacción con un determinado juguete".)

Los datos (100 en este caso) que habríamos obtenido, primeramente los podríamos describir haciendo uso de las técnicas vistas en el apartado 1.2: tablas de frecuencias, gráficas (histogramas, diagramas de cajas, etc.) y medidas descriptivas: media ($\overline{x}$), mediana, varianza (s^2), desviación típica (s), percentiles, etc. Pero como ya sabemos, no estamos interesados en esos 100 niños concretos, sino en todos los niños, toda la población. Para poder extraer conclusiones (inferir) acerca de la población (esto se verá en este curso), «necesitamos» asumir que nuestros datos provienen de una población que sigue un determinado modelo teórico (apartado 1.4, 1.5 y 1.6). A veces podría no asumirse un modelo paramétrico pero la estadística no paramétrica queda fuera de nuestro alcance. También existen tests para probar si nuestros datos provienen de un determinado modelo, que veremos en este curso.

Las conclusiones que obtendremos vendrán dadas en términos probabilísticos (por ejemplo, el intervalo de confianza al 95 % para μ es ...) y serán conclusiones sobre descriptores de la población (apartado 1.3): media (μ), varianza (σ^2), etc., que en realidad, muy difícilmente se conocen.

PARTE II
AMPLIACIÓN DE ESTADÍSTICA

Capítulo 2

Inferencia estadística. Estimación

Pienso que lo esencial, si quieres ser un buen estadístico a diferencia de ser un matemático, es hablar a la gente y averiguar lo que están haciendo y por qué lo están haciendo.

Florence Nightingale David

2.1. Introducción a la inferencia estadística

La inferencia estadística trata los métodos mediante los cuales podemos hacer inferencias (extraer determinadas conclusiones o generalizaciones) sobre una población, a partir de la información extraída de una muestra aleatoria de dicha población (como acabamos de repasar).

La inferencia estadística podría dividirse en dos áreas: la estimación y los contrastes de hipótesis. En este tema trataremos la estimación (aunque ya se trató en parte en el curso pasado) y en el siguiente, los contrastes de hipótesis.

- **Estimación**: Busca determinar el valor de una característica desconocida (parámetro) de la población.
- **Contraste de hipótesis**: Busca determinar si es aceptable que la característica estudiada cumpla cierta condición o comprobar una teoría o hipótesis sobre una población.

Veamos algunos ejemplos sencillos como ilustración:

Ejemplo 2.1 (Examen 3/9/2007): Se han propuesto diversos métodos para detectar si una persona no autorizada (un intruso) accede a una cuenta con el nombre de usuario y contraseña (robada o descifrada) correctas. Uno de ellos consiste en medir el tiempo entre las pulsaciones de las teclas. A un importante usuario autorizado (debidamente identificado) se le ha medido dicho tiempo, dando las 121 observaciones recogidas una media, $\bar{x} = 0.2$ segundos y una desviación típica (s) de 0.07 segundos. La media de dicha muestra puede emplearse para estimar la media de la población entera (todos los tiempos entre pulsaciones de dicho usuario), sin embargo, debe quedar claro que NO es

la media verdadera de la población. Emplearemos la distribución de muestreo de $\overline{X}$ para tener una idea de la exactitud de la estimación (**Problema de estimación**).

Ejemplo 2.2 (Examen 24/1/2007): A fin de verificar la adecuación de un sistema informático interactivo de venta de entradas de cine, se controla el tiempo de servicio de los usuarios. Para que este sistema sea satisfactorio, el tiempo de servicio medio por cliente no debe superar los 2 minutos. En efecto, los estudios realizados mostraron que un tiempo medio superior produce unas colas demasiado largas, y una espera que el usuario no está dispuesto a soportar; por lo tanto, el cine perderá clientes y dinero si el requisito mencionado no se satisface. Para controlar el tiempo de servicio, se observa una muestra aleatoria simple de 31 usuarios en uno de los cines de la cadena (en el *ABC vamo ar sine*), para saber si se debe o no proceder a la modificación del sistema informático de venta. El tiempo de servicio medio observado en la muestra es de 2.17 minutos y la desviación típica de 0.4 minutos. Para comprobar si se debe modificar el sistema informático actual, se plantearía la hipótesis de que el tiempo medio de servicio no supera los dos minutos, y tras las pruebas oportunas, dicha hipótesis podrá o no podrá ser rechazada. En este ejemplo no se pretende estimar un parámetro, sino decidir sobre una hipótesis. La teoría del muestreo también nos ayudará a determinar la exactitud de nuestra decisión (**Problema de contraste de hipótesis**).

2.2. Estimación

Distinguiremos dos tipos:

a) Estimación puntual

Se trata de estimar un parámetro poblacional mediante un número que lo aproxime. En el ejemplo 2.1, estimamos la media de la población (μ) con la media de una muestra ($\overline{x}$). Sin embargo, no podemos esperar que una estimación puntual coincida exactamente con el parámetro poblacional que pretende estimar, por ello en muchas ocasiones será preferible proporcionar un intervalo que contendrá al parámetro poblacional con un grado razonable de certidumbre.

b) Estimación por intervalos

Obtendremos intervalos, en los que confiamos que se encuentre el parámetro poblacional a estimar, por ejemplo la media poblacional μ. A estos intervalos se les conoce como intervalos de confianza para el parámetro al $(1 - \alpha) \cdot 100\,\%$, donde $1 - \alpha$ es el grado o nivel de confianza o también intervalos de confianza al nivel de significación α. (α estará entre 0 y 1, valores comunes son: 0.1, 0.05 y 0.01). Cuanto mayor sea $1 - \alpha$ (nivel de confianza), más amplio será el intervalo.

Formalmente, definimos un **intervalo de confianza** como un intervalo aleatorio cuyos puntos extremos T_1 y T_2 ($T_1 < T_2$) son funciones de la muestra aleatoria y tales que:

$$P(T_1(X_1, \ldots, X_n) \leq \theta \leq T_2(X_1, \ldots, X_n)) \geq 1 - \alpha,$$

donde al valor $1 - \alpha$ se le llama *nivel de confianza* y $0 < \alpha < 1$.

Para una realización de la muestra concreta, x_1, x_2, ..., x_n, obtenemos un intervalo numérico $(T_1(x_1, \ldots, x_n)$, $T_2(x_1, \ldots, x_n))$ que se llamará (abusando del lenguaje) también intervalo de confianza. Obviamente, en este caso no tiene sentido hablar de probabilidad, y por ello, diremos que tenemos una confianza de $100(1 - \alpha)\,\%$, en el sentido de que si tomásemos infinitas realizaciones, x_1, x_2, ..., x_n, de la muestra y construyésemos los correspondientes intervalos numéricos, el $100(1-\alpha)\,\%$ de estos intervalos contendrían el valor del parámetro y los restantes no. Debe recalcarse que es el intervalo el que es aleatorio, y no el parámetro.

¿Cuál es la interpretación de un intervalo de confianza?

Supongamos que construimos un intervalo de confianza al 95 % para μ, para una serie de muestras de una población Normal, cada una de ellas formada por, por ejemplo, 20 observaciones. Cada vez tendremos una media muestral ($\overline{x}$) diferente, mientras que μ no varía. Entonces, el 95 % de los intervalos que construyésemos contendrá a μ. Por supuesto, en un experimento concreto sólo disponemos de una muestra (formada por los 20 datos) y esperaremos "con confianza" que nuestra muestra sea una de las del 95 % (¡cuidado!: no tiene sentido hablar de la probabilidad de que μ esté en un intervalo, ya que aunque μ es desconocida, no es una variable aleatoria, sino entraríamos en el campo de la inferencia Bayesiana). Veámoslo gráficamente:

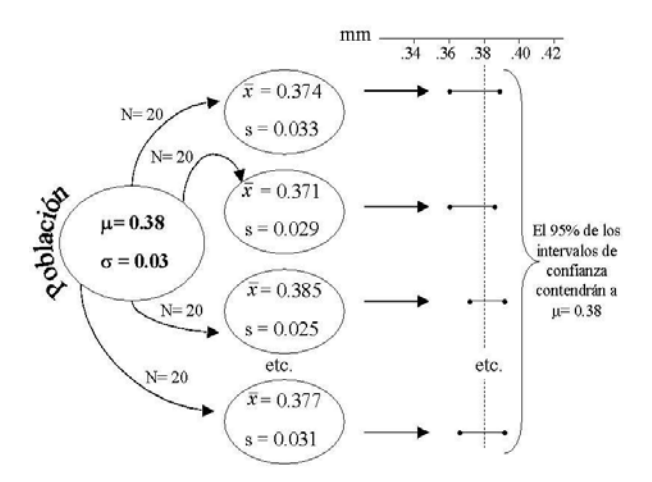


Figura 2.1: El 95 % de los intervalos de confianza contendrán a $\mu = 0.38$. El tamaño muestral considerado cada vez es 20

Si en lugar de 20, el tamaño muestral en cada muestra fuera 5, los intervalos

serán más grandes, pero nuevamente el 95 % de los intervalos de confianza contendrán a $\mu = 0.38$, según la siguiente gráfica.

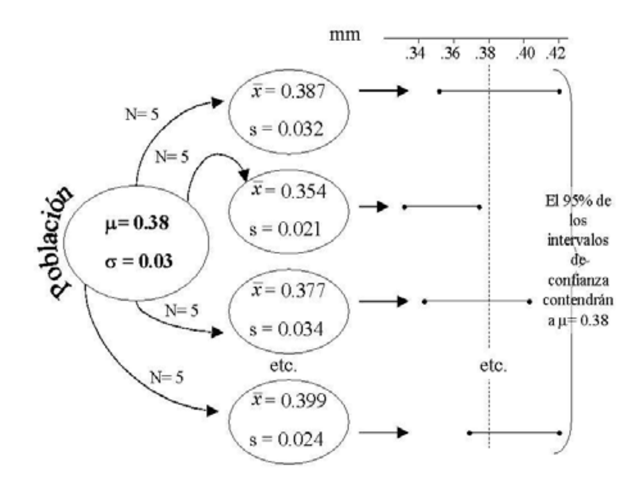


Figura 2.2: El 95 % de los intervalos de confianza contendrán a $\mu = 0.38$. El tamaño muestral considerado cada vez es 5

2.2.1. Estimación puntual

Se ha repasado en el apartado 1.7.1, aún así, veamos simplemente cómo estimar ciertos parámetros de determinadas distribuciones:

i) Estimador puntual de p, para una Binomial(n,p), n conocido:

$\hat{p} = \frac{X}{n}$ donde X es el número de éxitos que ocurren en las n observaciones.

ii) Estimador puntual de μ, para una Normal(μ,σ^2):

$\hat{\mu} = \overline{X}$.

Ejemplo 2.1 (continuación): $\hat{\mu} = \overline{x} = 0.2$ segundos. □

iii) Estimador puntual de σ^2, para una Normal(μ,σ^2):

$S^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N-1}$.

Ejemplo 2.1: Hacemos el muestreo y $s^2 = 0.07^2$ □

Si en lugar de haber dividido por $N - 1$, hubiésemos dividido por N, habríamos obtenido un estimador sesgado, es decir, $E(S^2) = \sigma^2$, mientras que $E(\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}) = (N - 1/N)\sigma^2$.

iv) Estimador puntual del parámetro λ de una Poisson:

$\hat{\lambda} = \overline{X}$.

2.3. Estimación por intervalos

A lo largo de este apartado N denotará el tamaño muestral y α el nivel de significación.

A continuación se examinarán algunos casos particulares de intervalos de confianza para los parámetros más importantes: medias, varianzas y proporciones. Aunque primero, veremos la construcción de un intervalo de confianza para un parámetro desconocido θ (en particular μ), usando un estadístico pivote, es decir, un estadístico cuya distribución en el muestreo no depende de θ. Para el resto de casos se haría análogamente:

A) Intervalo de confianza para μ, con σ^2 conocida:

Sea X_1, X_2, ..., X_N una muestra aleatoria de una población con media μ desconocida y σ^2 conocida. $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{N}}$ es aproximadamente N(0,1) si N es grande (por el teorema central del límite).

Por tanto, P($-z_{\alpha/2} \leq Z \leq z_{\alpha/2}$) = 1 - α, donde $z_{\alpha/2}$ es tal que P($Z \geq z_{\alpha/2}$) = $\alpha/2$.

Por ejemplo, para $\alpha = 0.05$:

$P(Z \geq 1.96) = 0.05/2 = 0.025$ y $P(-1.96 \leq Z \leq 1.96) = 0.95 \rightarrow$
$P(-1.96 \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{N}} \leq 1.96) = 0.95 \rightarrow$
$P(-1.96 \cdot \sigma/\sqrt{N} \leq \overline{X} - \mu \leq 1.96 \cdot \sigma/\sqrt{N}) = 0.95 \rightarrow$
$P(-1.96 \cdot \sigma/\sqrt{N} - \overline{X} \leq - \mu \leq 1.96 \cdot \sigma/\sqrt{N} - \overline{X}) = 0.95 \rightarrow$
$P(\overline{X} + 1.96 \cdot \sigma/\sqrt{N} \geq \mu \geq \overline{X} - 1.96 \cdot \sigma/\sqrt{N}) = 0.95 \rightarrow$

$P(\overline{X} - 1.96 \cdot \sigma/\sqrt{N} \leq \mu \leq \overline{X} + 1.96 \cdot \sigma/\sqrt{N}) = 0.95$.

$(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{N})}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}})$ con P(Z $\geq z_{\alpha/2}$) = $\alpha/2$, $Z \sim$ N(0,1)

B) Intervalo de confianza para μ, con σ^2 desconocida, para Normales:

$(\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{N}}, \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{N}})$ con P(T $\geq t_{\alpha/2}$) = $\alpha/2$, T es t-Student con $N-1$ grados de libertad

R: t.test(x, conf.level = 0.95, ...)

Nota: la distribución t se denomina t de Student por el seudónimo empleado por Gosset para publicar sus trabajos. Gosset trabajaba para la Guiness Brewers (la cervecería) en Irlanda, y debido a que su patrón desaprobaba la

publicación de investigaciones de un empleado, tuvo que publicar sus resultados bajo el seudónimo Student.

Ejemplo 2.3 (Examen 21/1/2009): Imagina que disponemos de un programa para simular el sistema y obtener los resultados (asumidos normales), en el contexto siguiente: un supermercado, para el que se plantean dos estrategias de distribución de cajas de pago y colas. Para cada una de las cuales, hemos simulado réplicas independientes y obtenido el tiempo medio que los clientes estarían haciendo cola:

Estrategia 1 (la actualmente en uso): 1.91, 1.82, 1.71, 1.83, 2.2, 2.4

Estrategia 2: 1.53, 1.66, 1.24, 2.34, 2

Centrándonos únicamente en la estrategia 2, encuentra el intervalo de confianza al 95 % para el tiempo medio de espera en cola.

Usaremos el caso B, ya que el enunciado nos dice que el tiempo de espera es Normal, y no conocemos σ^2: $(\overline{x} \pm t_{\alpha/2}\frac{s}{\sqrt{N}})$. A partir de los $N = 5$ datos del enunciado para la estrategia 2, obtenemos $\overline{x} = 1.754$ y $s = 0.426357$. A partir de las tablas (las tablas estaban en el libro de Gregori y Epifanio [34]), como $\alpha = 0.05$ (pues la confianza es 95 %) y los grados de libertad son N - 1 = 5 - 1 = 4, $t_{\alpha/2} = t_{0.05/2} = t_{0.025} = 2.776$. Por tanto, $(\overline{x} \pm t_{\alpha/2}\frac{s}{\sqrt{N}}) = (1.754 \pm 2.776\frac{0.426357}{\sqrt{5}}) = (1.754 \pm 0.529393) = (1.754 - 0.529393, 1.754 + 0.529393) = (1.22461, 2.28339)$

C) Intervalo de confianza para μ, con σ^2 desconocida y N grande ($N \geq 30$):

$(\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{N}}, \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{N}})$ con $P(Z \geq z_{\alpha/2}) = \alpha/2$, $Z \sim N(0,1)$

Observación: Aun cuando la normalidad no pueda suponerse, si deseamos obtener un intervalo de confianza para μ con la varianza desconocida, si la muestra es grande, podemos usar C). Si la muestra es pequeña, usaremos B) si la distribución es normal.

Fíjate que $z_{\alpha/2}$ cumple: $P(Z \geq z_{\alpha/2}) = \alpha/2$, $Z \sim N(0,1)$, es decir, la probabilidad que la variable Z sea mayor que $z_{\alpha/2}$ es $\alpha/2$.

A partir de la tabla de la Normal(0,1), podemos tener los valores críticos que más frecuentemente usaremos: $z_{0.1} = 1.28$, $z_{0.05} = 1.64$, $z_{0.025} = 1.96$, $z_{0.01} = 2.33$, $z_{0.005} = 2.57$.

Ejemplo 2.1: Construye un intervalo de confianza al 99 % para el tiempo medio entre pulsaciones de dicho usuario.

En el enunciado del ejemplo, no aparece señalado que el tiempo entre pulsaciones se distribuya normalmente, pero el tamaño muestral (121) es grande, así que usaremos el caso C, para obtener el intervalo de confianza. En este ejemplo, $\alpha = 0.01$, puesto que la confianza es 99 %, así que sustituyendo en el intervalo $(\overline{x} \pm z_{\alpha/2}\frac{s}{\sqrt{N}}$, tendríamos $(0.2 \pm 2.57\frac{0.07}{\sqrt{121}} = (0.18{,}0.22)$

Para determinar el tamaño muestral necesario para una precisión determinada, en el caso de la estimación de la media μ a partir de una muestra aleatoria simple, usaremos:

$$N = \left(\frac{z_{\alpha/2} \cdot \sigma}{Error}\right)^2$$

Cuando σ es desconocida, podemos recurrir a estudios previos o bien a la obtención de una muestra piloto previa, con la que estimaremos σ, mediante s.

Ejemplo 2.1: Si deseamos que el error en la estimación del tiempo medio anterior sea inferior a 0.01 segundos con una confianza del 95 %, y teniendo en cuenta que podríamos asumir $\sigma = 0.07$, ¿cuántas observaciones deberían recogerse como mínimo?

$$N = \left(\frac{z_{\alpha/2} \cdot \sigma}{Error}\right)^2 = \left(\frac{1.96 \cdot 0.07}{0.01}\right)^2.$$

Al menos, deberían recogerse 189 observaciones.

Aunque no entre en la materia del curso, no está de más conocer que, a veces, el interés no está en la estimación de parámetros, sino en *dónde caen las observaciones individuales*. Así pues, debemos distinguir entre intervalos de confianza e intervalos de tolerancia. Para una distribución Normal con media y varianza desconocidas, los límites de tolerancia están dados por $\overline{x} \pm ks$, donde k está determinado de modo que se pueda establecer con una confianza del $100(1 - \alpha)$ por ciento que los límites contienen al menos una proporción p de la población. En Montgomery y Runger [49] (por ejemplo), puedes encontrar las tablas que proporcionan k, con las que calcular estos intervalos de tolerancia, y más información sobre este punto.

A continuación, consideramos dos muestras aleatorias simples, X_1, X_2, ..., X_n e Y_1, Y_2, ..., Y_m obtenidas de dos poblaciones de interés X e Y, con el fin de construir intervalos que permitan comparar parámetros (seguimos con las medias) de X e Y.

Debemos diferenciar entre muestras independientes y apareadas. Hay veces que las muestras no son independientes, sino dependientes. Pueden ser apareadas (emparejadas), como es el caso de tener datos del tipo «antes» y «después» en el mismo individuo u objeto (unidad experimental), o bien si a cada objeto (u objetos emparejados) se le aplican dos métodos, o sea, que por cada unidad experimental tendremos dos observaciones, a diferencia de las muestras

independientes, donde las unidades experimentales son seleccionadas de forma independiente. En el caso de un muestreo apareado, conseguiremos grupos más homogéneos, reduciéndose la variabilidad experimental.

D) Intervalo de confianza para la diferencia de medias μ_1 - μ_2, con σ_1^2 y σ_2^2 conocidas, para muestras aleatorias independientes (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

$(\overline{x}_1 - \overline{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}})$ con $P(Z \geq z_{\alpha/2}) = \alpha/2$, $Z \sim$ N(0,1)

E) Intervalo de confianza para la diferencia de medias μ_1 - μ_2, con σ_1^2 y σ_2^2 desconocidas, para muestras aleatorias independientes y tamaños muestrales grandes (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

$(\overline{x}_1 - \overline{x}_2 \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}})$ con $P(Z \geq z_{\alpha/2}) = \alpha/2$, $Z \sim$ N(0,1)

Para el caso de una diferencia entre dos medias, la interpretación del intervalo de confianza puede extenderse a una comparación de las dos medias. De esta manera, por ejemplo, si tenemos gran confianza de que una diferencia μ_1 - μ_2 es positiva, realmente inferiremos que $\mu_1 > \mu_2$ con poco riesgo de caer en un error. Por tanto, en la interpretación de los intervalos de confianza para diferencia de medias nos fijaremos si el cero pertenece al intervalo o no. Piensa que si son iguales, su resta vale cero, con lo cual si cero no está incluido en el intervalo, indicaría que las medias poblacionales son diferentes.

F) Intervalo de confianza para la diferencia de medias μ_1 - μ_2 de poblaciones normales independientes, con varianzas poblacionales desconocidas pero iguales ($\sigma_1^2 = \sigma_2^2$) (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

$(\overline{x}_1 - \overline{x}_2 \pm t_{\alpha/2}\sqrt{\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{N_1+N_2-2}}\sqrt{\frac{N_1+N_2}{N_1 N_2}})$ con $P(T \geq t_{\alpha/2}) = \alpha/2$, T es t-Student con $N_1 + N_2 - 2$ grados de libertad

R: t.test(x,y, var.equal = TRUE, conf.level = 0.95, ...)

Ejemplo 2.3: Calcula el intervalo de confianza de la diferencia de tiempos medios de espera al 95 % para ambas estrategias, asumiendo igualdad de varianzas (lo comprobaremos en un apartado posterior). ¿Podemos suponer igualdad de medias poblacionales?

Según el enunciado, podemos asumir normalidad e igualdad de varianzas, además como los tamaños muestrales son pequeños ($N_1 = 6$ y $N_2 = 5$), y las muestras son independientes, usaremos el caso F. Con la calculadora, pode-

mos obtener los estadísticos muestrales que necesitamos (media y desviaciones típicas de cada muestra), e introducirlos en la fórmula:

$(\overline{x}_1$ - $\overline{x}_2 \pm t_{\alpha/2}\sqrt{\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{N_1+N_2-2}}\sqrt{\frac{N_1+N_2}{N_1N_2}}) = (1.97833$ - $1.754 \pm t_{0.05/2}\sqrt{\frac{(6-1)0.264833^2+(5-1)0.426357^2}{6+5-2}}\sqrt{\frac{6+5}{6\cdot5}}) = (0.22433 \pm 0.47403) =$ (-0.25,0.698).

Puesto $0 \in$ (-0.25,0.698), no podemos afirmar que exista diferencia entre los tiempos de espera medios (μ_1 y μ_2) de ambas estrategias, al 95 % de confianza.

G) Intervalo de confianza para la diferencia de medias μ_1 - μ_2 de poblaciones normales independientes, con varianzas poblacionales σ_1^2, σ_2^2 desconocidas y desiguales (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

$(\overline{x}_1$ - $\overline{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}})$ con $P(T \geq t_{\alpha/2}) = \alpha/2$, T es t-Student con grados de libertad $\frac{(\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2})^2}{\frac{(s_1^2/N_1)^2}{N_1-1}+\frac{(s_2^2/N_2)^2}{N_2-1}}$

R: t.test(x,y, conf.level = 0.95, ...)

Ejemplo 2.1: Un día se graban los siguientes tiempos entre pulsaciones para una entrada correcta de dicho usuario en segundos: 0.38, 0.31, 0.24, 0.2, 0.31, 0.34, 0.42, 0.09, 0.18, 0.46, 0.21. Asumiendo normalidad, y diferentes varianzas (que ya es algo sospechoso), por lo que a los tiempos medios entre pulsaciones se refiere, usa el intervalo de confianza apropiado, para determinar si hay evidencia de que un intruso ha accedido a la cuenta de dicho usuario (con $\alpha = 0.05$).

Construiremos el intervalo de confianza de diferencia de medias para muestras independientes con el caso G, siguiendo lo que podemos leer en el enunciado.

En este problema, los grados de libertad de la t-Student son:

$$\frac{(\frac{0.07^2}{121}+\frac{0.112^2}{11})^2}{\frac{(0.07^2/121)^2}{121-1}+\frac{(0.112^2/11)^2}{11-1}} \simeq 11$$

$(\overline{x}_1$ - $\overline{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}}) = (0.2$ - $0.285 \pm t_{0.05/2}\sqrt{\frac{0.07^2}{121}+\frac{0.112^2}{11}}) =$ (-0.085455 $\pm 2.2 \cdot 0.034447) =$ (-0.16124, -0.00967)

Como $0 \notin$ (-0.16124, -0.00967), sí que hay diferencia entre las medias, sí que habría accedido un intruso.

H) Intervalo de confianza para la diferencia de medias μ_D, para muestras apareadas, con diferencia normal.

$(\overline{d} \pm t_{\alpha/2}\, \frac{s_d}{\sqrt{N}})$ donde $\overline{d}$ es la media de las diferencias y s_d es la desviación típica de las diferencias. Además, $P(T \geq t_{\alpha/2}) = \alpha/2$, T es t-Student con N - 1 grados de libertad, N es el número de objetos (parejas) de que disponemos

R: t.test(x,y, paired = TRUE, conf.level = 0.95, ...)

Ejemplo 2.4 (Examen 1/9/2008): Este verano ha aparecido en los medios de comunicación un estudio realizado sobre los beneficios del *Brain Training*. En concreto, durante quince semanas un grupo de personas jugó diariamente durante un periodo corto de tiempo. A continuación, aparece un subconjunto adaptado de estos datos, con los tiempos en segundos (que asumiremos normal) para el ejercicio de agilidad mental, Cálculo 20, en la primera y última semana para 7 individuos:

	Semana 1	Semana 15
Sujeto 1	60	31
Sujeto 2	50	25
Sujeto 3	54	20
Sujeto 4	74	35
Sujeto 5	58	24
Sujeto 6	65	23
Sujeto 7	57	22

Construyamos el intervalo de confianza de la diferencia de medias al 95 %, para comprobar si existe diferencia entre ambas.

Obviamente las muestras no son independientes, son apareadas, ya que medimos al mismo individuo, sin jugar y tras jugar varias semanas con el *Brain Training*. Por ello, como además nos dicen que son Normales, usaremos el caso H.

Lo primero será obtener, las diferencias de la Semana 1 - Semana 15: 29 25 34 39 34 42 35, a partir de las cuales calculamos $\overline{d} = 34$ y $s_d = 5.71548$, que introducimos en la fórmula: $(\overline{d} \pm t_{\alpha/2}\, \frac{s_d}{\sqrt{N}}) = (34 \pm t_{0.05/2}\, \frac{5.71548}{\sqrt{7}}) = (34 \pm 5.28595) = (28.7141, 39.2859)$

Claramente $0 \notin (28.7141, 39.2859)$, en consecuencia hay diferencia entre los tiempos medios entre antes y después del uso del *Brain Training*.

A continuación, nos interesamos por otros parámetros:

I) Intervalo de confianza para σ^2 en una población normal:

$(\frac{(N-1)s^2}{\chi^2_{\alpha/2}}, \frac{(N-1)s^2}{\chi^2_{1-\alpha/2}})$ con $P(\chi^2 > \chi^2_{\alpha/2}) = \alpha/2$, χ^2 es chi- cuadrado con $N-1$ grados de libertad.

Ejemplo 2.4: Construye un intervalo de confianza de 99 % para la desviación típica del Tiempo de la Semana 15.

El intervalo de confianza al 99 % para la varianza poblacional σ^2 es:

$(\frac{(7-1)5.34522^2}{\chi^2_{0.01/2}},\frac{(7-1)5.34522^2}{\chi^2_{1-0.01/2}}) = (\frac{(7-1)5.34522^2}{18.54},\frac{(7-1)5.34522^2}{0.6757}) = (9.2464{,}253.70469)$

Como buscamos el de la desviación típica σ:

$(\sqrt{9.2464},\sqrt{253.70469}) = (3.041{,}15.928)$

J) Intervalo de confianza para el cociente σ_1^2/σ_2^2 de varianzas de dos poblaciones normales independientes:

$(\frac{s_1^2}{s_2^2}\frac{1}{F_{\alpha/2}},\frac{s_1^2}{s_2^2}\frac{1}{F_{1-\alpha/2}})$ donde P(F > $F_{\alpha/2}$) = $\alpha/2$ y F es F de Snedecor con $(N_1 - 1,\ N_2 - 1)$ grados de libertad

R: var.test(x, y, conf.level = 0.95, ...)
En la interpretación de los intervalos de confianza para cociente de varianzas nos fijaremos si el uno pertenece al intervalo o no. Piensa que si son iguales, su cociente vale uno, con lo cual si uno no está incluido en el intervalo, indicaría que las varianzas poblacionales son diferentes.

Ejemplo 2.3: Construye un intervalo de confianza al 95 % para el cociente de varianzas de ambas estrategias. ¿Fue apropiado suponer igualdad de varianzas?

El intervalo es:

$(\frac{s_1^2}{s_2^2}\frac{1}{F_{\alpha/2}},\frac{s_1^2}{s_2^2}\frac{1}{F_{1-\alpha/2}}) = (\frac{0.264^2}{0.426^2}\frac{1}{F_{0.05/2}},\frac{0.264^2}{0.426^2}\frac{1}{F_{1-0.05/2}}) = (\frac{0.264^2}{0.426^2}\frac{1}{9.36},\frac{0.264^2}{0.426^2}\frac{1}{1/7.39}) =$
$(0.041{,}2.85)$

Es claro que $1 \in (0.041{,}2.85)$, así que sí que fue apropiado suponer igualdad de varianzas.

K) Intervalo de confianza para una proporción p (de una Binomial) cuando N es grande y la proporción no es cercana a cero o uno:

$(\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{N}})$, donde P($Z > z_{\alpha/2}$) = $\alpha/2$ $Z \sim$ N(0,1) y $\hat{p}$ = X /N, $\hat{q}$ = 1 - $\hat{p}$, X = número de éxitos

R: prop.test(x, n, conf.level = 0.95)

Nota: aunque en prácticas usemos esta función que lleva el R en la base (en la librería stats), esta función no devuelve el intervalo anterior que es el que suele aparecer en los libros de texto, sino el intervalo basado en el estadístico

score sin corrección de continuidad, y que según [1] sería preferible. Si quisiéramos obtener el intervalo que calcularemos en teoría, que es más sencillo de calcular a mano, tendríamos que usar la función binconf de la librería Hmisc con la opción "asymptotic".

Ejemplo 2.5 (Examen 9/9/2005): Se toma una muestra de estudiantes universitarios de informática y se les pregunta por su sistema operativo favorito, como resultado se obtiene que de 200 encuestados 30 prefieren el *Macrochof*. Si p es la proporción de 'preferencia del *Macrochof*' entre los estudiantes de informática, calcula un intervalo de confianza al 95 % para p.

Usando el caso K, obtendríamos: $(\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{N}}) = (30/200 \pm 1.96\sqrt{\frac{0.15\cdot 0.85}{200}})$ $= (0.15 \pm 0.0495) = (0.1005{,}0.1995)$

La magnitud del error que cometemos al emplear X/N como estimador de p, viene dada por: E = Error = $z_{\alpha/2}\sqrt{\frac{p(1-p)}{N}}$. Esta fórmula nos puede servir para determinar el tamaño muestral (que obviamente tendrá que ser un entero) necesario para alcanzar un grado de precisión deseado.

$$N = \hat{p}(1-\hat{p}) \cdot (\frac{z_{\alpha/2}}{E})^2$$

Si no dispusiésemos de información acerca del valor de p, por ejemplo en base a una muestra piloto, podríamos considerar el peor caso, es decir, con la que obtendríamos la máxima N, cuando la proporción valiera 0.5:

$$N = p(1-p) \cdot (\frac{z_{\alpha/2}}{E})^2 \leq \frac{1}{4}\,(\frac{z_{\alpha/2}}{E})^2$$

Una vez obtenidos los N datos, tendremos un $(1-\alpha)100\,\%$ de confianza que el error no excederá E.

Ejemplo 2.5: Queremos estimar la proporción de 'preferencia del *Macrochof*', y deseamos estar al menos 95 % seguros que el error es como mucho de 0.03. ¿Cómo ha de ser de grande la muestra?, si:

a) no tenemos idea de cuál pueda ser la proporción real.

$$N = \frac{1}{4}\,(\frac{z_{\alpha/2}}{E})^2 = \frac{1}{4}\,(\frac{1.96}{0.03})^2 = 1067.1 \uparrow 1068$$

b) usamos la proporción anterior (0.15) como una muestra preliminar que nos proporciona una estimación preliminar.

$$N = \hat{p}(1-\hat{p}) \cdot (\frac{z_{\alpha/2}}{E})^2 = 0.15(1-0.15) \cdot (\frac{1.96}{0.03})^2 = 544.2 \uparrow 545$$

L) Intervalo de confianza para una proporción p, si ésta es muy cercana a cero:

$(0, \frac{1}{2N}\chi^2_\alpha)$ con $P(\chi^2 > \chi^2_\alpha) = \alpha$, χ^2 es Chi-cuadrado con $2(X+1)$ grados de libertad, X = número de éxitos

M) Intervalo de confianza para la diferencia de dos proporciones, con N_1 y N_2 grandes (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

$(\hat{p_1}$ - $\hat{p_2} \pm z_{\alpha/2}\sqrt{\frac{\hat{p_1}\hat{q_1}}{N_1} + \frac{\hat{p_2}\hat{q_2}}{N_2}})$, donde P($Z > z_{\alpha/2}) = \alpha/2$ $Z \sim$ N(0,1), $\hat{p_1} = X_1/N_1$, $\hat{q_1} = 1$ - $\hat{p_1}$, X_1 = número de éxitos en las N_1 pruebas y $\hat{p_2} = X_2\,/N_2$, $\hat{q_2} = 1$ - $\hat{p_2}$, X_2 = número de éxitos en las N_2 pruebas

R: prop.test(x, n, conf.level = 0.95)

En la interpretación de los intervalos de confianza para diferencia de proporciones nos fijaremos si el cero pertenece al intervalo o no. Piensa que si son iguales, su resta vale cero, con lo cual si cero no está incluido en el intervalo, indicaría que las proporciones poblacionales son diferentes.

Ejemplo 2.5: Además de la muestra anterior de la universidad A, se toma una muestra de estudiantes universitarios de informática de otra universidad, la B, y se les pregunta por su sistema operativo favorito, como resultado que en la universidad B prefieren *Macrohof* 60 de los 300 encuestados. Desea determinarse si las preferencias por *Macrohof* difieren en ambas universidades, así que determina el intervalo de confianza al 95 % para la diferencia de proporciones.

Usaremos el caso M:

$(\hat{p_1}$ - $\hat{p_2} \pm z_{\alpha/2}\sqrt{\frac{\hat{p_1}\hat{q_1}}{N_1} + \frac{\hat{p_2}\hat{q_2}}{N_2}}) = (0.15$ - $0.2 \pm 1.96\sqrt{\frac{0.15\cdot 0.85}{200} + \frac{0.2\cdot 0.8}{300}}) =$ (-0.117,0.017)

Como $0 \in$ (-0.117,0.017), no podemos afirmar que haya diferencia entre ambas proporciones.

Capítulo 3

Contrastes de hipótesis

El pensamiento estadístico será algún día tan necesario para el ciudadano competente como la habilidad de leer y escribir.

Herbert George Wells

3.1. Introducción

Hay muchos problemas (como fue el ejemplo 2.2) en los que más que estimar el valor de un parámetro, debemos decidir si un enunciado referente a un parámetro es cierto o falso, o sea, debemos probar una hipótesis sobre un (o más) parámetro(s), o bien, comprobar si una teoría sobre la población es verosímil dados los datos muestrales. En ellos, el objetivo de la experimentación es avalar o rechazar las afirmaciones realizadas y no la estimación de sus valores reales. Entramos en una parte importantísima de la inferencia estadística: las *pruebas o contrastes de hipótesis*, donde se formula una hipótesis, se experimenta y se juzga si los resultados apoyan estadísticamente la hipótesis de partida. No obstante, como nos movemos en condiciones de incertidumbre, la decisión final se realizará en términos probabilísticos.

Contraste de hipótesis: Es un método numérico para comprobar una teoría o hipótesis sobre una población.

En primer lugar, se comenzará con una introducción a los contrastes de hipótesis, tratando los conceptos básicos, tras lo cual se verán los contrastes relativos a la media, varianza y proporciones a partir de una y dos muestras, como en el tema anterior se hizo para intervalos de confianza. La última parte del tema se dedicará a algunos contrastes no paramétricos. Es decir, en la primera parte del tema se supondrá conocida la distribución teórica de la(s) variable(s) de interés, excepto en los valores de parámetros que la determinan, este tipo de hipótesis se denominarán *hipótesis paramétricas*, distinguiendo entre *hipótesis simples* e *hipótesis compuestas* según si especifican un único valor o un intervalo de valores para el parámetro.

En todo contraste de hipótesis nos encontramos con una hipótesis nula (H_0) y una hipótesis alternativa (H_1 o H_A). Cuando la meta de un experimento sea

establecer una afirmación, ésta se convertirá en la hipótesis alternativa y su negación será la hipótesis nula. La hipótesis nula se supone cierta hasta que los datos indiquen lo contrario, por tanto, la que se ha de demostrar que es cierta es la hipótesis alternativa, H_1. Podríamos plantear un símil con los juicios. En principio, se parte de la hipótesis nula de que un acusado es inocente hasta que se demuestre su culpabilidad. El fiscal es el que debe demostrar la culpabilidad y no será hallado culpable a menos que la hipótesis nula de su inocencia sea claramente rechazada. Con lo cual, si es hallado no culpable, no implica que el acusado haya demostrado su inocencia, sino que sólo implica que no se ha demostrado su culpabilidad.

En definitiva, se denominará *hipótesis nula*, H_0, a aquella que se contrasta, y es aquella que se mantiene a menos que los datos indiquen su falsedad. La hipótesis nula nunca se considera probada. Por ello, si el experimentador quiere respaldar con contundencia un determinado argumento es debido a que éste no puede ser asumido gratuitamente y, por tanto, sólo podrá ser defendido a través del rechazo del argumento contrario (el establecido en H_0). El rechazo de H_0 implica aceptar como correcta una hipótesis complementaria, la *hipótesis alternativa*, H_1.

Ejemplo 2.2 (continuación): A fin de verificar la adecuación de un sistema informático interactivo de venta de entradas de cine, se controla el tiempo de servicio de los usuarios. Para que este sistema sea satisfactorio, el tiempo de servicio medio por cliente no debe superar los 2 minutos. En efecto, los estudios realizados mostraron que un tiempo medio superior produce unas colas demasiado largas, y una espera que el usuario no está dispuesto a soportar; por lo tanto, el cine perderá clientes y dinero si el requisito mencionado no se satisface. Para controlar el tiempo de servicio, se observa una muestra aleatoria simple de 31 usuarios en uno de los cines de la cadena (en el *"ABC vamo ar sine"*), para saber si se debe o no proceder a la modificación del sistema informático de venta. El tiempo de servicio medio observado en la muestra es de $\overline{x} = 2.17$ minutos y la desviación típica de $s = 0.4$ minutos. Denotemos por μ el tiempo de servicio medio, así que $\overline{x}$ es un valor estimado de μ. Para comprobar si se debe modificar el sistema informático actual, se plantearía el siguiente contraste:

$$\begin{cases} H_0: & \mu \leq \mu_0 = 2 \text{ (o simplemente } \mu = \mu_0 = 2) \\ H_1: & \mu > \mu_0 = 2 \end{cases}$$

Planteamos dicho contraste ya que la hipótesis que queremos demostrar es la necesidad de modificar el sistema. Implícitamente se está suponiendo que el sistema actual es bueno y que los gastos de modificación son lo suficiente importantes como para necesitar una justificación.

Nota: Aunque en este caso sea algo rebuscado, si hubiésemos partido de que el sistema existente es malo y los gastos del cambio no fueran importantes, entonces habría que demostrar que el existente es satisfactorio y plantearíamos:

$$\begin{cases} H_0: & \mu \geq \mu_0 = 2 \\ H_1: & \mu < \mu_0 = 2 \end{cases}$$

En este tema, para contrastes paramétricos estudiaremos únicamente contrastes con hipótesis nula simple, $H_0 : \theta = \theta_0$, puesto que los contrastes con hipótesis nula compuesta del tipo $H_0 : \theta \geq \theta_0$ o $\theta \leq \theta_0$ y H_1 unilateral, $H_1 : \theta < \theta_0$ o $H_1 : \theta > \theta_0$ respectivamente, equivaldrán al contraste simple $H_0 : \theta = \theta_0$ frente al unilateral. Intuitivamente, podemos pensar en el ejemplo anterior que si verdaderamente $\mu < 2$, más difícil (y cuanto más menor que 2, más difícil) será que los datos respalden la hipótesis alternativa $\mu > 2$, de esta manera nos protegemos contra la peor posibilidad, el peor escenario, que sería cuando $\mu = 2$.

En consecuencia se considerarán únicamente estos tres casos, donde a) es un contraste bilateral mientras que b) y c) son contrastes unilaterales.

a) $\begin{array}{l} H_0 : \theta = \theta_0 \\ \\ H_1 : \theta \neq \theta_0 \end{array}$ b) $\begin{array}{l} H_0 : \theta \geq \theta_0 \ (\theta = \theta_0) \\ \\ H_1 : \theta < \theta_0 \end{array}$ c) $\begin{array}{l} H_0 : \theta \leq \theta_0 \ (\theta = \theta_0) \\ \\ H_1 : \theta > \theta_0 \end{array}$

Planteado el contraste, pasamos a explicar la metodología de los contrastes, es decir, tendremos que establecer algún criterio estadístico que permita decidir hasta qué punto los datos están o no de acuerdo con la hipótesis nula. En un contraste de hipótesis, se analizan los datos observados para ver si permiten rechazar la H_0, comprobando si estos datos tienen una probabilidad de aparecer lo suficientemente pequeña cuando la hipótesis nula es cierta.

Por esto, es necesario definir una medida de discrepancia entre los datos muestrales y la hipótesis nula. Para contrastes paramétricos, la discrepancia puede expresarse como una función del parámetro especificado por H_0, θ_0, y el valor estimado en la muestra, $\hat{\theta}$, $d(\theta_0; \hat{\theta})$, que llamaremos *estadístico de contraste*, y de la que conoceremos su distribución cuando H_0 sea cierta. Así, si H_0 es cierta, se conocerá la probabilidad de superar el valor que el estadístico de contraste haya tomado para una muestra concreta. Si esta probabilidad es «grande», no hay razones para sospechar que la hipótesis nula sea falsa, pero si es «pequeña», ello sólo puede atribuirse a dos causas: o bien la aleatoriedad de la muestra o bien que la distribución teórica supuesta para el estadístico de contraste es errónea, lo cual nos conduciría a haber asumido una hipótesis nula falsa. Por tanto, definir un contraste de significación requiere: una medida de discrepancia y una regla para juzgar qué discrepancias son «demasiado» grandes.

El método tradicional de realizar un contraste es dividir el rango de discrepancias que pueden observarse cuando H_0 es cierta en dos regiones: una *región de aceptación* de H_0 y otra de *rechazo o región crítica*. Se consideran discrepancias «demasiado» grandes aquellas que tienen una probabilidad, α, pequeña de ocurrir si H_0 es cierta, por ello, si rechazamos H_0 cuando ocurre

una discrepancia de probabilidad α, este número, que llamaremos *nivel de significación*, podemos interpretarlo como la probabilidad que estamos dispuestos a asumir de rechazar H_0 cuando es cierta, o sea, $\alpha = \mathrm{P}(\text{rechazar } H_0 \mid H_0 \text{ cierta})$. Fijado α (habitualmente 0.1, 0.05 o 0.01), la región crítica se determina a partir de la distribución del estadístico de contraste cuando H_0 es cierta, y dependerá del tipo de hipótesis alternativa.

Para una hipótesis nula simple, $H_0 : \theta = \theta_0$, y las distintas hipótesis alternativas que consideramos, las regiones críticas tendrán la siguiente forma, para el nivel de significación α, que también pueden apreciarse en la figura 3.1:

1. $H_1 : \theta \neq \theta_0 \qquad RR_\alpha = (-\infty, d_{1-\alpha/2}) \cup (d_{\alpha/2}, \infty)$
2. $H_1 : \theta > \theta_0 \qquad RR_\alpha = (d_\alpha, \infty)$
3. $H_1 : \theta < \theta_0 \qquad RR_\alpha = (-\infty, d_{1-\alpha})$

donde D es el estadístico de contraste y se ha denotado por d_α el valor tal que $\alpha = \mathrm{P}(D \geq d_\alpha \mid \theta = \theta_0)$, que denominaremos valor crítico. En consecuencia, la decisión tomada sobre un contraste deberá acompañarse del nivel de significación α prefijado, pues en realidad, todo contraste puede ser (o no) significativo dependiendo del nivel de significación.

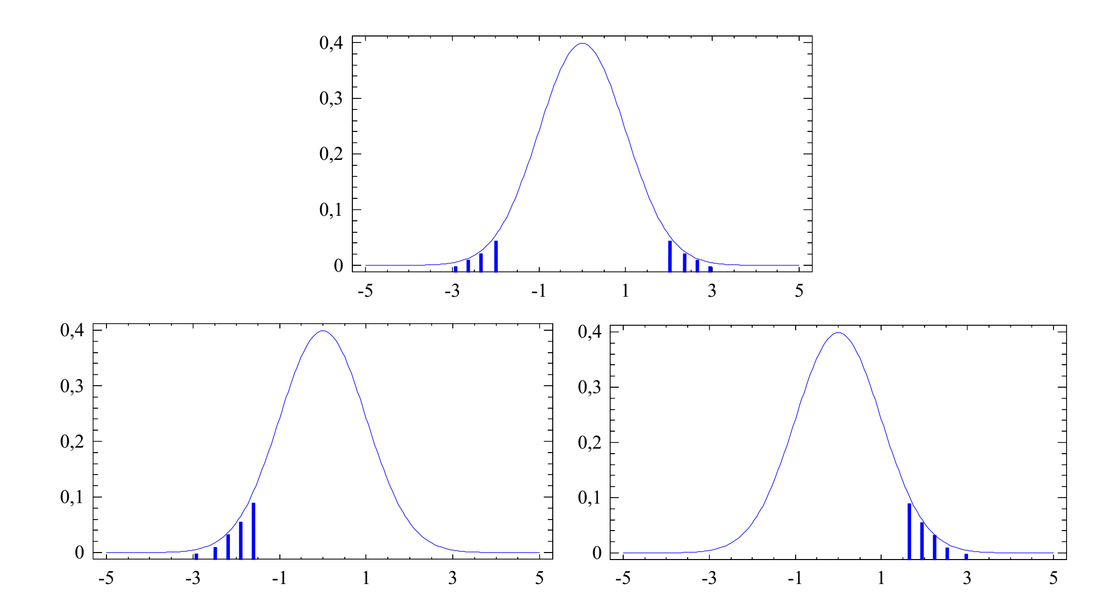


Figura 3.1: Regiones críticas para las siguientes hipótesis alternativas respectivamente: $H_1 : \theta \neq \theta_0$; $H_1 : \theta < \theta_0$; $H_1 : \theta > \theta_0$

Por este motivo, se define el *nivel crítico* o *p-valor* como la probabilidad de obtener una discrepancia mayor o igual (en relación con el distanciamiento de H_0 en la dirección de H_1) que la observada en la muestra, cuando H_0 es cierta. Este concepto es de suma importancia, pues además los paquetes estadísticos

expresan los resultados en términos de p-valores. El p-valor sólo puede calcularse una vez tomada la muestra. El p-valor puede interpretarse como un nivel mínimo de significación, de manera que niveles de significación α mayores o iguales que el p-valor conducirán a rechazar H_0, mientras que valores de α menores, conducirán a no rechazarla.

Puesto que la decisión que se tome en favor de una u otra hipótesis estará basada en la discrepancia observada entre la hipótesis nula y la información suministrada por la muestra, existirá la posibilidad de cometer dos tipos de errores:

Error de tipo I: se produce cuando H_0 es cierta pero se rechaza. La probabilidad de cometerlo se designa por α.
Error de tipo II: se produce cuando H_0 es falsa pero se "acepta". La probabilidad de incurrir en él se designa por β.

Se resume en la siguiente tabla:

	"Aceptar" H_0	Rechazar H_0
H_0 cierta	Decisión correcta	Error I
H_0 falsa	Error II	Decisión correcta

Llamaremos *función o curva característica* a la función que asigna a cada posible valor del parámetro θ, la probabilidad de aceptar H_0 cuando θ es cierto, o sea, $\beta(\theta) = \text{P(aceptar } H_0|\theta)$. Para $\theta = \theta_0$, $\beta(\theta_0) = \text{P(aceptar } H_0|\theta_0)$ $= 1 - \alpha$, mientras que para otros valores proporciona la probabilidad de cometer un error de tipo II. Debe señalarse que cuanto menor sea α mayor será $\beta(\theta)$ y al revés, la única forma de disminuir la probabilidad de ambos errores simultáneamente es aumentar el tamaño muestral. En ocasiones también se utiliza la *curva de potencia*, que indica la probabilidad de rechazar H_0 para cada valor del parámetro: $\text{Potencia}(\theta) = \text{P(rechazar } H_0|\theta)$. Si la potencia permanece siempre muy próxima a 1, entonces se dice que el estadístico de contraste es muy potente para contrastar H_0, ya que en ese caso las muestras resultarán, con alta probabilidad, incompatibles con H_0 cuando H_1 sea cierta. Por tanto, la potencia de un contraste puede interpretarse como su *sensibilidad* o capacidad para detectar una hipótesis alternativa.

Recopilando todo lo considerado hasta ahora, pueden establecerse los siguientes pasos para contrastar una hipótesis estadística:

1) Formular una hipótesis nula y alternativa apropiada.

2) Especificar la probabilidad de error de tipo I, según cómo de importante se considere una decisión errónea en favor de la hipótesis alternativa.

3) Elegir un estadístico de contraste D adecuado, así como su distribución.

4) Evaluar el estadístico de contraste D, para la muestra $x_1, x_2, \ldots, x_n$, para obtener el valor $d = D(x_1, x_2, \ldots, x_n)$.

5) Determinación de la región crítica.

6) Decisión: rechazar H_0 si el valor observado d, pertenece a la región crítica, sino no rechazar H_0.

Obviamente, así planteado, no controlaríamos el riesgo de cometer un error de tipo II, en caso de desear controlarlo, deberíamos determinar cuál es el primer valor de la hipótesis alternativa (θ_1) que, de ser correcto, deseamos detectar, además de especificar el tamaño del error de tipo II ($\beta(\theta_1)$) que estamos dispuestos a asumir. A partir de las probabilidades α y β, calcularíamos el tamaño muestral adecuado para garantizar ambas probabilidades de error.

Ejemplo 2.2: Utiliza el contraste adecuado para comprobar si se debe modificar el sistema informático actual, a nivel de significación de 0.05.

Vamos a ir paso a paso, para resolverlo:

1) Formular una hipótesis nula y alternativa apropiada:

Ya lo habíamos hecho:

$$\begin{cases} H_0: & \mu \leq \mu_0 = 2 \text{ (o simplemente } \mu = \mu_0 = 2) \\ H_1: & \mu > \mu_0 = 2 \end{cases}$$

2) Especificar la probabilidad de error de tipo I e identificar los datos con los que contamos.

$\mu_0 = 2$, $s = 0.4$, $\overline{x} = 2.17$, $N = 31$, $\alpha = 0.05$.

3) Elegir un estadístico de contraste adecuado, así como su distribución (véase la página 66):

Como es un contraste sobre una media y N es grande, elegimos el caso A:

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{31}} \sim N(0,1)$$

4) Cálculo del valor observado del estadístico de contraste según los datos observados:

$$z = \frac{2.17 - 2}{0.4/\sqrt{31}} = 2.3663$$

5) Determinación de la región crítica y el/los valor/es crítico/s:

Región crítica: son los valores del estadístico de contraste que nos conducen a rechazar la hipótesis nula.

Región de aceptación: son los valores del estadístico de contraste que nos llevan a "aceptar" = *no rechazar* la hipótesis nula.

Valor crítico: valor/es que separan la región crítica de la de aceptación.

Como la hipótesis alternativa es $\mu > 2$, elegimos la región crítica correspondiente a >: $(z_\alpha, +\infty)$

Como $\alpha = 0.05$ y el estadístico sigue una N(0,1), $z_\alpha = z_{0.05} = 1.64$ y la región crítica quedaría: $(1.64, \infty)$.

6) Decisión: rechazar H_0 si el valor observado pertenece a la región crítica, sino no rechazar H_0.

$2.3663 \in (1.64, \infty) \rightarrow$ Rechazamos H_0, nos quedamos con H_1. La media es mayor que 2, y por tanto, el sistema debe modificarse.

En los software estadísticos, para resolución de contrastes, se nos devuelve habitualmente el **p-valor**. Además de calcular la región crítica puede calcularse el $p - valor$ (asociado a nuestros datos). Recordemos que el $p - valor$ es el menor valor de α que nos conduciría a rechazar H_0. Un $p - valor$ se determina como la probabilidad de que el estadístico de contraste pertenezca a la región crítica cuando el valor observado se considera valor crítico. Valores pequeños del $p - valor$ (por ejemplo menor que 0.05) nos llevan a rechazar H_0. Si α es menor que el $p - valor$, no rechazamos H_0. En cambio, si α es mayor que el $p - valor$, rechazamos H_0.

Ejemplo 2.2: Podemos obtener en este ejemplo fácilmente el p-valor (en otros ejemplos, su obtención manual no será tan sencilla, ya que no dispondremos de esa información en las tablas, pero el ordenador siempre nos podrá sacar del apuro).

$$p - valor = P(Z > 2.366) = 1 - 0.9911 = 0.0089$$

Observación: Existe una relación entre los intervalos de confianza y los contrastes de hipótesis. Los intervalos de confianza (bilaterales) vistos en el tema anterior (exceptuando el L)) nos dan la región de aceptación de contrastes bilaterales al $100 \cdot (1 - \alpha)\,\%$ y por tanto, H_0 no será rechazada al nivel α si θ_0 pertenece al intervalo. O sea, intervalo de confianza $(1 - \alpha)$ = conjunto de hipótesis aceptables a nivel α. Por ejemplo, para el caso de la media de una población Normal, se acepta al nivel α la hipótesis $\mu = \mu_0$ cuando el intervalo de confianza 1 - α construido para μ incluye a μ_0 y viceversa.

Ejemplo 1.9: Consideremos ventiladores de ordenador, que en condiciones normales, tienen una vida distribuida normalmente con media 15.100 horas.

Se introducen ciertos cambios en el proceso de fabricación que pueden afectar a la media pero no a la variabilidad. Para contrastar si estos cambios han producido efectos, tomamos una muestra de 4 ventiladores cuyas vidas resultan ser (en horas): 15010, 14750, 14826, 14953. ¿Hay evidencia de un efecto sobre la media?

$\overline{x}$ = 14884.75, s = 118.258, intervalo de confianza para μ al 95 %: ($\overline{x} \pm t_{\alpha/2}\frac{s}{\sqrt{N)}}$) = (14884.75 ± 3.182$\frac{118.258}{\sqrt{4)}}$) = (14696.6,15072.9), como 15100 $\notin$ (14696.6,15072.9) → sí que habrá afectado a la media.

Ambos procedimientos (contrastes e intervalos) deben considerarse complementarios.

Hemos de distinguir también entre diferencias estadísticamente significativas y la significatividad práctica. Por ejemplo, si tomamos una muestra muy grande y tratamos de contrastar si la media es μ_0, nos puede ocurrir que observemos una diferencia significativa, rechazando que la media sea μ_0, cuando en realidad, la media sea $\mu_0 + 0.00001$, una diferencia que puede no ser importante a nivel práctico. Así, además del contraste de hipótesis, es conveniente realizar una estimación de los parámetros y un análisis de la potencia para evaluar la capacidad de discriminación.

3.2. Contrastes paramétricos: medias, varianzas y proporciones

Ahora trataremos algunos contrastes paramétricos de interés práctico para una y dos muestras: medias, varianzas y proporciones. Para cada caso, se expondrá el estadístico de contraste correspondiente y la región de rechazo.

A) Contraste de hipótesis para la media μ, con N grande ($N \geq 30$):

$$Z \approx \frac{\bar{X} - \mu_0}{S/\sqrt{N}} \sim N(0,1) \qquad \begin{cases} H_0 : \mu = \mu_0 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$\mu < \mu_0$	$(-\infty, -z_\alpha)$
$\mu \neq \mu_0$	$(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$
$\mu > \mu_0$	(z_α, ∞)

Ejemplo 2.2: se ha resuelto en el apartado anterior.

B) Contraste de hipótesis para la media μ en una población Normal con σ^2 desconocida:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{N}} \sim t_{N-1} \qquad \begin{cases} H_0 : \mu = \mu_0 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$\mu < \mu_0$	$(-\infty, -t_\alpha)$
$\mu \neq \mu_0$	$(-\infty, -t_{\alpha/2}) \cup (t_{\alpha/2}, \infty)$
$\mu > \mu_0$	(t_α, ∞)

R: t.test(x, alternative = c("two.sided", "less", "greater"), mu = 0, conf.level = 0.95, ...)

C) Contraste para la diferencia de medias μ_1 - μ_2, con σ_1^2 y σ_2^2 desconocidas, para muestras aleatorias independientes y tamaños muestrales grandes (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

$$Z \approx \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{s_1^2/N_1 + s_2^2/N_2}} \sim N(0,1) \qquad \begin{cases} H_0 : \mu_1 - \mu_2 = \Delta_0 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$\mu_1 - \mu_2 < \Delta_0$	$(-\infty, -z_\alpha)$
$\mu_1 - \mu_2 \neq \Delta_0$	$(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$
$\mu_1 - \mu_2 > \Delta_0$	(z_α, ∞)

Ejemplo 2.2: En dicho cine de la cadena, se instala de prueba un nuevo sistema informático, por comprobar si mejora el anterior y recomendar su cambio en el resto de cines de la cadena con largas colas. Esta vez, se observan 41 usuarios, con media y desviación típica de 1.9 y 0.35, respectivamente. Usa el contraste que determine si el sistema nuevo mejora el anterior, en cuanto a los tiempos de servicio medios se refiere (con $\alpha = 0.05$).

Puesto que hemos de realizar un contraste para dos medias, las muestras son independientes y los tamaños muestrales son grandes (31 y 41), usaremos el caso C. Si μ_1 es el tiempo medio de servicio con el sistema anterior, y μ_2 con el nuevo, el sistema nuevo mejorará el anterior, si disminuye el tiempo de servicio medio, o sea, si $\mu_1 > \mu_2$, o lo que es lo mismo, $\mu_1 - \mu_2 > 0$.

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 > 0 \end{cases}$$

Sustituyamos nuestros datos en el estadístico de contraste:

$$z = \frac{2.17 - 1.9 - 0}{\sqrt{0.4^2/31 + 0.35^2/41}} = 2.99095$$

$2.991 \in$ Región Crítica $= (z_\alpha, \infty) = (z_{0.05}, \infty) = (1.64, \infty)$, por tanto, rechazo H_0, el tiempo de servicio medio con el anterior es mayor que con el nuevo, es decir, el nuevo mejora el anterior.

D) Contraste para la diferencia de medias μ_1 - μ_2 de poblaciones normales independientes, con varianzas poblacionales desconocidas pero iguales ($\sigma_1^2 = \sigma_2^2$) (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

$$T = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{N_1+N_2-2}}} \sqrt{\frac{N_1 \cdot N_2}{N_1 + N_2}} \sim t_{N_1+N_2-2} \qquad \begin{cases} H_0 : \mu_1 - \mu_2 = \Delta_0 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$\mu_1 - \mu_2 < \Delta_0$	$(-\infty, -t_\alpha)$
$\mu_1 - \mu_2 \neq \Delta_0$	$(-\infty, -t_{\alpha/2}) \cup (t_{\alpha/2}, \infty)$
$\mu_1 - \mu_2 > \Delta_0$	(t_α, ∞)

R: t.test(x,y, alternative = c("two.sided", "less", "greater"), mu = 0, var.equal = TRUE, conf.level = 0.95, ...)

Ejemplo 3.1 (Examen 26/1/2005): En una multinacional que se dedica a la venta de baterías para portátiles, se consideran dos modelos. El departamento de ingeniería ha realizado pruebas de duración para los modelos bajo condiciones de uso y recarga similares, que se recogen a continuación:

Modelo viejo	8500	9500	9600	8400	9400	8300
Modelo nuevo	10000	9800	10300	9900	10200	

¿Puede concluirse al nivel $\alpha = 0.05$ que la duración media del modelo nuevo es 800 horas superior que para el modelo viejo? Utiliza el contraste adecuado para responder la pregunta anterior. Asume normalidad y que sus varianzas no difieren.

Es claro que el caso a utilizar es el D. Si μ_1 es la duración media para el modelo viejo y μ_2 para el nuevo, la duración media del modelo nuevo es 800 horas superior que para el modelo viejo cuando $\mu_2 > \mu_1 + 800$, o sea, cuando $-\mu_1 + \mu_2 > 800$, es decir, cambiando el signo, $\mu_1 - \mu_2 < -800$. En caso de dudas, puede ayudarte el ponerte ejemplos numéricos.

$$\begin{cases} H_0 : \mu_1 - \mu_2 = -800 \\ H_1 : \mu_1 - \mu_2 < -800 \end{cases}$$

$$t = \frac{8950 - 10040 - (-800)}{\sqrt{\frac{5 \cdot 371000 + 4 \cdot 43000}{6+5-2}}} \sqrt{\frac{6 \cdot 5}{6 + 5}} = -1.009$$

-1.009 $\notin$ Región crítica = $(-\infty, -t_\alpha) = (-\infty, -t_{0.05})$ = (g.l. son 6 + 5 - 2 = 9) = $(-\infty, -1.833)$ En consecuencia, no rechazo H_0, no tenemos pruebas para afirmar que el modelo nuevo sea 800 horas superior al viejo, en cuanto a la duración media se refiere.

E) Contraste para la diferencia de medias μ_1 - μ_2 de poblaciones normales independientes, con varianzas poblacionales σ_1^2, σ_2^2 desconocidas y desiguales (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

$$T = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{s_1^2/N_1 + s_2^2/N_2}} \sim t_{\text{g.l.}}$$

$$g.l. = \frac{(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2})^2}{\frac{(s_1^2/N_1)^2}{N_1-1} + \frac{(s_2^2/N_2)^2}{N_2-1}} \qquad \begin{cases} H_0 : \mu_1 - \mu_2 = \Delta_0 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$\mu_1 - \mu_2 < \Delta_0$	$(-\infty, -t_\alpha)$
$\mu_1 - \mu_2 \neq \Delta_0$	$(-\infty, -t_{\alpha/2}) \cup (t_{\alpha/2}, \infty)$
$\mu_1 - \mu_2 > \Delta_0$	(t_α, ∞)

R: t.test(x,y, alternative = c("two.sided", "less", "greater"), mu = 0, conf.level = 0.95, ...)

F) Contraste para la diferencia de medias μ_1 - μ_2 para muestras apareadas, cuya diferencia es normal: $\bar{D}$ y S_D son la media y desviación típica de las diferencias

$$T = \frac{\bar{D} - \Delta_0}{S_D/\sqrt{N}} \sim t_{N-1} \qquad \begin{cases} H_0 : \mu_D = \Delta_0 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$\mu_D < \Delta_0$	$(-\infty, -t_\alpha)$
$\mu_D \neq \Delta_0$	$(-\infty, -t_{\alpha/2}) \cup (t_{\alpha/2}, \infty)$
$\mu_D > \Delta_0$	(t_α, ∞)

R t.test(x,y, alternative = c("two.sided", "less", "greater"), mu = 0, paired=TRUE, conf.level = 0.95, ...)

Ejemplo 3.2 (Examen 6/2/2004): Se quiere comparar la rapidez de dos modelos de impresora A y B. Los de la compañía A sostienen que su modelo es más de 5 segundos más rápido que para el modelo B de los rivales, respecto a tiempos medios de impresión. Se mide el tiempo de impresión de los dos

modelos (que consideraremos normal) sobre una serie de 8 plantillas estándar y los resultados aparecen en la tabla siguiente:

	Tiempo para A	Tiempo para B
Plantilla 1	20	26
Plantilla 2	25	29
Plantilla 3	22	27
Plantilla 4	23	28
Plantilla 5	19	29
Plantilla 6	21	30
Plantilla 7	18	25
Plantilla 8	20	26

Utiliza el contraste adecuado para comprobar si este estudio confirma la afirmación de la compañía A a nivel de significación de 0.05.

Claramente por cada plantilla contamos con una pareja de datos, los tiempos para el A y B, así que usaremos el caso F. Calculamos la muestra de la variable diferencia, D = Tiempo con A - Tiempo con B, para nuestros datos, que proporciona $\overline{d}$ = -6.5 y s_d = 2.07 que incluiremos en el estadístico de contraste. Fíjate que si el modelo A es más de 5 segundos más rápido que el modelo B, se traduce en que $\mu_D < -5$, ya que más rápido, equivale a menos tiempo.

$$\begin{cases} H_0 : \mu_D = -5 \\ H_1 : \mu_D < -5 \end{cases}$$

$$t = \frac{-6.5 - (-5)}{2.07/\sqrt{8}} = -2.049$$

-2.049 $\in$ Región Crítica $= (-\infty, -t_\alpha) = (-\infty, -t_{0.05})$ = (g.l. = N - 1 = 7, ya que hay N = 8 parejas) $= (-\infty, -1.895)$

Así que rechazo H_0, sí que tenía razón la compañía A, en media la impresora A es más de 5 segundos más rápida que la B.

G) Contraste para σ^2 en una población normal:

$$\chi_0^2 = \frac{(N-1)S^2}{\sigma_0^2} \sim \chi_{N-1}^2 \qquad \begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$\sigma^2 < \sigma_0^2$	$(0, \chi_{1-\alpha}^2)$
$\sigma^2 \neq \sigma_0^2$	$(0, \chi_{1-\alpha/2}^2) \cup (\chi_{\alpha/2}^2, \infty)$
$\sigma^2 > \sigma_0^2$	(χ_α^2, ∞)

H) Contraste para el cociente σ_1^2/σ_2^2 de varianzas de dos poblaciones normales independientes:

$$F = \frac{S_1^2}{S_2^2} \sim F_{(N_1-1,N_2-1)} \quad \begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$\sigma_1^2 < \sigma_2^2$	$(0, F_{1-\alpha}) = (0, \frac{1}{F_\alpha^{(N_2-1,N_1-1)}})$
$\sigma_1^2 \neq \sigma_2^2$	$(0, F_{1-\alpha/2}) \cup (F_{\alpha/2}, \infty)$
$\sigma_1^2 > \sigma_2^2$	(F_α, ∞)

R: var.test(x, y, ratio = 1, alternative = c("two.sided", "less", "greater"), conf.level = 0.95, ...)

Ejemplo 2.1: ¿Puede concluirse al nivel $\alpha = 0.01$ que la varianza del tiempo entre pulsaciones para el usuario autorizado es menor que para la entrada recogida posteriormente?

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 < \sigma_2^2 \end{cases}$$

$$f = \frac{0.07^2}{0.112282^2} = 0.388665$$

$0.388665 \in$ Región Crítica $= (0, F_{1-\alpha}) = (0, \frac{1}{F_\alpha^{(N_2-1,N_1-1)}}) = (0, F_{1-0.01}) = (0, \frac{1}{F_{0.01}^{(11-1,121-1)}}) = (0, \frac{1}{F_{0.01}^{(11-1,121-1)}}) = (0, \frac{1}{2.559}) = (0,\ 0.390778)$ Rechazo H_0, la varianza del tiempo entre pulsaciones para el usuario autorizado sí es menor que para la entrada recogida posteriormente, aunque ciertamente estamos en la frontera.

I) Contraste para una proporción p (de una Binomial) cuando N es grande y la proporción no es cercana a cero ni a uno:
$\hat{p} = X/N$ (X = número de éxitos en las N pruebas), q_0 = 1 - p_0

$$Z \approx \frac{\hat{p} - p_0}{\sqrt{p_0 q_0/N}} \sim N(0,1) \quad \begin{cases} H_0 : p = p_0 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$p < p_0$	$(-\infty, -z_\alpha)$
$p \neq p_0$	$(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$
$p > p_0$	(z_α, ∞)

R: prop.test(x, n, p= NULL,alternative = c("two.sided", "less", "greater"),conf.level = 0.95)

J) Contraste para la diferencia de dos proporciones, con N_1 y N_2 grandes (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):
$\hat{p_1} = X_1/N_1$ (X_1 = número de éxitos en las N_1 pruebas), $\hat{p_2} = X_2/N_2$ (X_2 = número de éxitos en las N_2 pruebas), $\hat{p} = (X_1 + X_2)/(N_1 + N_2)$

$$Z \approx \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})(1/N_1 + 1/N_2)}} \sim N(0,1) \quad \begin{cases} H_0 : p_1 = p_2 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$p_1 < p_2$	$(-\infty, -z_\alpha)$
$p_1 \neq p_2$	$(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$
$p_1 > p_2$	(z_α, ∞)

R: prop.test(x, n, p= NULL,alternative = c("two.sided", "less", "greater"),conf.level = 0.95)

Ejemplo 3.3 (Examen 2/2/2008): Se desea construir un gráfico de control (lo veremos en el tema próximo) para controlar un proceso que fabrica diodos para un circuito impreso. Se tienen 19 muestras, cada una formada por 50 diodos. El número de diodos defectuosos en cada una de las muestras aparece a continuación: 4 5 3 1 4 5 7 5 6 5 1 3 1 2 3 5 4 6 2. Al cabo del tiempo, tras un reajuste, el operario sospecha que la proporción de diodos defectuosos puede haber disminuido, así que toma aleatoriamente 200 diodos de la cadena de producción y comprueba que 10 son defectuosos. Utiliza el contraste adecuado para comprobar la sospecha del operario, a nivel de significación de 0.05.

Puesto que estamos tratando con dos proporciones usaremos el caso J. Si llamamos p_1 y p_2 a la proporción de defectuosos antes y después del reajuste, respectivamente:

$$\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 > p_2 \end{cases}$$

Antes del reajuste, vemos que hay 72 defectuosos de $19 \cdot 50 = 950$ diodos, o sea, $\hat{p_1} = 0.07579$. Mientras que después, $\hat{p_2} = 10/\ 200 = 0.05$. Por otro lado, $\hat{p} = (X_1 + X_2)/(N_1 + N_2) = (72 + 10)/950 + 200) = 0.0713$

$$z = \frac{0.07579 - 0.05}{\sqrt{0.0713(1 - 0.0713)(1/950 + 1/200)}} = 1.288$$

$1.288 \notin$ Región crítica $= (z_\alpha, \infty) = (z_{0.05}, \infty) = (1.64, \infty)$ Así que, no rechazo H_0, no tengo razones para afirmar que $p_1 > p_2$, por tanto, no se ha confirmado la sospecha de que la proporción de diodos defectuosos haya disminuido tras el ajuste.

3.3. Test de la χ^2

Hasta ahora nos hemos centrado en la inferencia paramétrica, donde son claves las siguientes tres hipótesis: independencia de los datos, ajuste a la distribución especificada y homogeneidad, es decir, que no tengamos heterogeneidad: muestras de distintas distribuciones.

Por un lado debemos buscar procedimientos para evaluar el cumplimiento de dichas hipótesis, y por otro lado, cuando éstas no se cumplan, debemos buscar procedimientos alternativos a los vistos anteriormente. De todo esto, se encarga la inferencia no paramétrica.

Nosotros veremos únicamente los contrastes de bondad de ajuste y el análisis de tablas de contingencia. Si necesitarais ampliar este punto, el libro [11] es muy recomendable, ya que cuenta con una introducción a la inferencia no paramétrica muy clara, tratando: contrastes de localización (test de los signos, test de Wilcoxon de los rangos signados), contrastes de independencia (contrastes basados en rachas, contraste de Ljung-Box), contrastes de homogeneidad (contrastes de valores atípicos, contraste de Wald-Wolfowitz basado en rachas, contraste de suma de rangos de Wilcoxon y el de la U de Mann-Whitney, contraste de Kolmogorov-Smirnov para dos muestras).

En este apartado, veremos el test chi-cuadrado, χ^2, que puede adoptar dos formas que nos permitirán contrastar la bondad de ajuste y la independencia u homogeneidad en tablas de contingencia, como veremos a continuación.

Una **prueba de bondad de ajuste** se emplea para decidir cuando un conjunto datos se puede considerar que ha sido obtenido de una población con una distribución de probabilidad dada.

K) Prueba de la bondad de ajuste con la χ^2:

El objetivo de este contraste es aclarar si es cierta la hipótesis nula H_0 de que una variable sigue una distribución teórica determinada. Por ello, se tratará de ver si las frecuencias de las observaciones se ajustan bien con la distribución.

El contraste *ji-cuadrado de Pearson* es válido para todo tipo de distribuciones, discretas y continuas, si agrupamos las observaciones en un cierto número no demasiado pequeño k de intervalos. Con él podemos contrastar dos tipos de hipótesis nula, especificando completamente la distribución o especificando simplemente la forma pero no los parámetros, que se estiman a partir de los datos.

Para esta prueba, las observaciones de la muestra aleatoria de tamaño N de la población cuya distribución de probabilidad es desconocida se ordenan en un histograma de frecuencia, con k intervalos de clase. Denotaremos por o_i

la frecuencia observada en el intervalo de clase i. Calcularemos la frecuencia esperada, e_i, para el intervalo i-ésimo, a partir de la distribución de probabilidad hipotética.

El estadístico que usaremos es:

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}, \qquad (3.1)$$

que sigue aproximadamente una distribución χ^2 con $k-1$ grados de libertad, siempre que la distribución especificada sea la correcta. Sin embargo, es usual aplicar el test aun en casos en los que la distribución de la variable no está totalmente especificada, sino que depende de algún parámetro que, en consecuencia, deberá ser estimado (por ejemplo, el caso en que se suponga que la variable en concreto sigue una distribución de Poisson y falta por especificar su parámetro λ). En estos casos la distribución aproximada del test ha de ser corregida para incorporar esta información pasando a ser una χ^2 con $k-r-1$ grados de libertad, siendo r el número de parámetros estimados por máxima verosimilitud.

La hipótesis nula de que la distribución de la población es la distribución hipotética se rechazará si el valor calculado del estadístico anterior χ_0^2 es mayor que χ_α^2, o sea, la región crítica (a nivel α) es: (χ_α^2, ∞).

R: chisq.test(x, p = rep(1/length(x), length(x)))

Una limitación bastante recomendable en la práctica es la de no llevar a cabo el contraste cuando la frecuencia esperada de alguna clase sea menor que 5, para evitar problemas de mala aproximación de la distribución usada a la verdadera distribución. Entonces, en los casos en los que esta condición falle, podríamos agrupar varios valores adyacentes hasta que se cumpla la restricción.

Veamos primero un ejemplo (ejemplo 3.4) en el que la distribución viene completamente determinada, y posteriormente otro (ejemplo 3.5) en el que la distribución depende de uno o más parámetros desconocidos.

Ejemplo 3.4 (Examen 26/1/2005): Estamos interesados en comprobar la perfección de un dado cúbico (un dado normal de 6 caras), es decir, en comprobar si los resultados se distribuyen uniformemente. Con los resultados obtenidos en 60 lanzamientos del dado, decide si se distribuirían uniformemente usando $\alpha = 0.05$:

Resultado	1	2	3	4	5	6
Frecuencia	15	9	7	13	12	4

Planteamos el contraste de bondad de ajuste:

$$\begin{cases} H_0: & \text{Resultado al lanzar el dado es uniforme} \\ H_1: & \text{Resultado al lanzar el dado no es uniforme} \end{cases}$$

Construimos una tabla con las frecuencias observadas y esperadas (fíjate que en caso de seguir la uniforme, los resultados son equiprobables):

Resultado	Probabilidad del resultado p_i	Frecuencia esperada e_i	Frecuencia observada o_i
1	1/6	$1/6 \cdot 60 = 10$	15
2	1/6	10	9
3	1/6	10	7
4	1/6	10	13
5	1/6	10	12
6	1/6	10	4

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \frac{(15-10)^2}{10} + \frac{(9-10)^2}{10} +$$

$$\frac{(7-10)^2}{10} + \frac{(13-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(4-10)^2}{10} = 8.4$$

8.4 $\notin$ Región crítica = $(\chi^2_\alpha, \infty) = (\chi^2_{0.05}, \infty)$ = (g.l. = 6 - 1 = 5) = (11.07, ∞) Por tanto, no rechazo H_0, no tenemos pruebas suficientes para afirmar que no sea una uniforme.

Ejemplo 3.5 (Examen 3/9/2007): Se ha hecho un seguimiento durante una serie de días del número de mensajes spam al día que un cierto usuario recibe en su correo electrónico. En base a dichos datos, que se recogen en la tabla siguiente, decide si se ajustaría a una distribución de Poisson, considerando α = 0.05.

Número spam diario	0	1	2	3	4	5	≥ 6
Frecuencia observada	35	42	55	40	15	10	3

Planteamos el contraste de bondad de ajuste:

$$\begin{cases} H_0: & \text{X= “Número de spam diario” es Poisson} \\ H_1: & \text{X= “Número de spam diario” no es Poisson} \end{cases}$$

Para construir la tabla con las frecuencias observadas y esperadas, necesitamos previamente estimar el valor de la λ de la Poisson, mediante la media:

$$\hat{\lambda} = \overline{x} = (35{\cdot}0+42{\cdot}1+55{\cdot}2+40{\cdot}3+15{\cdot}4+10{\cdot}5+3{\cdot}6)/(35+42+55+40+15+10+3)$$

$$= (42 + 110 + 120 + 60 + 50 + 18)/200 = 400/200 = 2$$

Calculamos la probabilidad de que tome cada valor y la frecuencia esperada correspondiente. Recuerda que la función de probabilidad de la Poisson(λ) es: $P(X=x)=\frac{e^{-\lambda}\lambda^x}{x!}$

$$P(X=0)=\frac{e^{-2}2^0}{0!}=0.135335\rightarrow e_0=0.135335\cdot 200=27.067$$

$$P(X=1)=\frac{e^{-2}2^1}{1!}=0.270671\rightarrow e_1=0.270671\cdot 200=54.1342$$

$$P(X=2)=\frac{e^{-2}2^2}{2!}=0.270671\rightarrow e_2=0.270671\cdot 200=54.1342$$

$$P(X=3)=\frac{e^{-2}2^3}{3!}=0.180447\rightarrow e_3=0.180447\cdot 200=36.0894$$

$$P(X=4)=\frac{e^{-2}2^4}{4!}=0.0902235\rightarrow e_4=0.0902235\cdot 200=18.0447$$

$$P(X=5)=\frac{e^{-2}2^5}{5!}=0.0360894\rightarrow e_5=0.0360894\cdot 200=7.21788$$

No hace falta que calculemos $P(X\geq 6)$, ya que como la frecuencia esperada de cada categoría debe ser mayor que 5, y como la frecuencia esperada acumulada hasta el momento es: 27.067 + 54.1342 + 54.1342 + 36.0894 + 18.0447 + 7.21788 = 196.68738, y el total era 200, sólo tendríamos disponible 200 - 196.68738 = 3.31262 que es menor que 5. En consecuencia, agrupamos las dos últimas categorías, para disponer de una frecuencia esperada superior a 5.

X	Probabilidad de categoría p_i	Frecuencia esperada e_i	Frecuencia observada o_i
0	0.14	27.07	35
1	0.27	54.13	42
2	0.27	54.13	55
3	0.18	36.09	40
4	0.09	18.04	15
≥ 5	1-(0.14+0.27+0.27+0.18+0.09)	7.21+3.31=10.53	10+3=13

$$\chi_0^2=\sum_{i=1}^{k}\frac{(o_i-e_i)^2}{e_i}=\frac{(35-27.07)^2}{27.07}+\frac{(42-54.13)^2}{54.13}+$$

$$\frac{(55-54.13)^2}{54.13}+\frac{(40-36.09)^2}{36.09}+\frac{(15-18.04)^2}{18.04}+\frac{(13-10.53)^2}{10.53}=6.6$$

Al sacar los grados de libertad, debemos recordar que hemos estimado un parámetro, por tanto, tendremos 6 categorías (tras reagrupar) - 1 - 1 (por estimar λ) = 4 grados de libertad.

$6.6 \notin$ Región crítica = $(\chi^2_\alpha, \infty) = (\chi^2_{0.05}, \infty)$ = (g.l. = 6 - 1 - 1 = 4) = (9.49, ∞) Por tanto, no rechazo H_0, no tenemos pruebas suficientes para afirmar que no sea una Poisson.

Como ya hemos dicho, el contraste χ^2, se usa no sólo para variables discretas o cualitativas, sino incluso con variables continuas. En este caso, dicha variable ha de ser agrupada en intervalos. Obviamente, el resultado del test dependerá de cómo construyamos estos intervalos. En el siguiente apartado (3.4), veremos que en el caso de variables continuas, podremos usar otros tests.

L) Pruebas con tablas de contingencia:

Sea la tabla de contingencia siguiente:

X \ Y	y_1	...	y_j	...	y_c	Total
x_1	o_{11}	...	o_{1j}	...	o_{1c}	$T_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_i	o_{i1}	...	o_{ij}	...	o_{ic}	$T_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_r	o_{r1}	...	o_{rj}	...	o_{rc}	$T_{r.}$
Total	$T_{.1}$	...	$T_{.j}$	...	$T_{.c}$	T

donde $T_{i.}$ es el total de observaciones de la fila i-ésima, $T_{.j}$ es el total de observaciones de la columna j-ésima y T es el total de observaciones.

Una tabla de contingencia como la anterior puede surgir en dos contextos diferentes:

a) Una muestra y dos variables (X e Y) cada una de ellas con r y c valores. En este caso podría interesarnos **contrastar la hipótesis de independencia de las dos variables**.

$$\begin{cases} H_0: & \text{Las dos variables son independientes} \\ H_1: & \text{Las dos variables son dependientes (asociadas)} \end{cases}$$

Nota: El curso pasado ya se vio el concepto de independencia con detalle, en el capítulo 6 del libro [34].

Si las variables X e Y son independientes, $p_{ij} = P(X = i, Y = j) = P(X = i) \cdot P(Y = j) = p_{i.} \cdot p_{.j}$, para $i = 1, ..., r$, y $j = 1, ..., c$, por tanto, el contraste sería:

$$\begin{cases} H_0: & p_{ij} = p_{i.} \cdot p_{.j} \ \forall \ i,j \\ H_1: & p_{ij} \neq p_{i.} \cdot p_{.j} \text{ para algún } i,j \end{cases}$$

Ejemplo 2.5: En una universidad C, los 200 encuestados aleatoriamente entre los estudiantes de informática, pertenecían a tres carreras distintas: ITIG (Ingeniería Técnica en Informática de Gestión), ITIS (Ingeniería Técnica en Informática de Sistemas) e II (Ingeniería Informática). En la encuesta tenían dos opciones de sistema operativo (SO) favorito a elegir: *Macrohof* y *Pingüino*, de forma que los datos desglosados aparecen en la tabla siguiente:

	Carrera		
SO	ITIG	ITIS	II
Macrochof	30	25	15
Pingüino	50	55	25

Nos interesa saber a un nivel de significación del 5 % si la preferencia por un SO u otro es independiente de la carrera.

b) c muestras independientes de c poblaciones y una variable observada con r categorías, es decir, X_1, X_2, ..., X_c, cada una puede tomar r valores. En este caso, podría interesarnos contrastar la hipótesis de que todas las distribuciones de donde se seleccionan las c muestras son semejantes, es decir, que **las c distribuciones son homogéneas**. Por ejemplo, cuando existen sólo dos categorías, tales como éxito y fracaso, defectuoso y no defectuoso, etc., entonces la prueba de homogeneidad es en realidad una prueba sobre la igualdad de c parámetros binomiales.

La H_0, por tanto, quedaría de la siguiente forma:

$$\begin{cases} H_0: & p_{11} = p_{12} = ... = p_{1c} \\ & p_{21} = p_{22} = ... = p_{2c} \\ & \vdots \quad \vdots \quad \quad \vdots \\ & p_{r1} = p_{r2} = ... = p_{rc} \end{cases}$$

Las proporciones de las poblaciones son constantes en cualquier categoría, mientras que H_1 afirma que al menos en una de las categorías no todas las probabilidades serían iguales.

Ejemplo 3.6 (Examen 25/1/2006): Consideremos 3 proveedores chinos de reproductores de MP3, *Chin Lu*, *Chin Ga* y *Chin Na*; 200 MP3 de cada proveedor son aleatoriamente seleccionadas para examinarlas obteniéndose los siguientes resultados:

Calidad	Proveedor *Chin Lu*	*Chin Ga*	*Chin Na*
Defectuoso	6	4	16
No defectuoso	194	196	184

¿La proporción de defectuosos es la misma para los 3 proveedores, o sea, son homogéneas las poblaciones? (considera $\alpha = 0.01$).

Para ambos casos, a) y b), el cálculo del estadístico de contraste es el mismo, aunque la forma de plantear H_0 y de enunciar las conclusiones sean distintas.

El estadístico que usaremos es:

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}, \qquad (3.2)$$

siendo $e_{ij} = T_{i.} \cdot T_{.j}$ / T

Bajo H_0, sigue aproximadamente una distribución χ^2 con $(r-1)\cdot(c-1)$ grados de libertad. La región crítica (a nivel α) es: (χ^2_α, ∞). Para que la aproximación sea correcta, todas las e_{ij} deben ser al menos 5.

R: chisq.test(x), siendo x una matriz.

Ejemplo 2.5: Primero calculamos χ_0^2.

SO	Carrera ITIG	ITIS	II	Total
Macrochof	30	25	15	30+25+15 = 70
Pingüino	50	55	25	50+55+25 = 130
Total	30+50 = 80	25+55 = 80	15+25 = 40	200

$$\chi_0^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = \frac{(30-70\cdot 80/200)^2}{70\cdot 80/200} + \frac{(25-70\cdot 80/200)^2}{70\cdot 80/200} +$$

$$\frac{(15-70\cdot 40/200)^2}{70\cdot 40/200} + \frac{(50-130\cdot 80/200)^2}{130\cdot 80/200} + \frac{(55-130\cdot 80/200)^2}{130\cdot 80/200} +$$

$$\frac{(25-130\cdot 40/200)^2}{130\cdot 40/200} = 0.82$$

Como hay $r = 2$ filas y $c = 3$ columnas, los grados de libertad serán $(r-1)\cdot(c-1)$ = (2 - 1) · (3 -1) = 2 grados de libertad.

$0.82 \notin$ Región Crítica $= (\chi^2_\alpha, \infty) = (\chi^2_{0.05}, \infty) = (5.99, \infty)$ No rechazo H_0, no tengo pruebas para afirmar que no sean independientes.

Ejemplo 3.6:

	Proveedor			
Calidad	*Chin Lu*	*Chin Ga*	*Chin Na*	Total
Defectuoso	6	4	16	26
No defectuoso	194	196	184	574
Total	200	200	200	600

De forma análoga, en este ejemplo calculamos χ^2_0 que es igual a 9.97.

$9.97 \in$ Región Crítica $= (\chi^2_\alpha, \infty) = (\chi^2_{0.01}, \infty) =$ (con (2 - 1) $\cdot$ (3 -1) = 2 grados de libertad) $= (9.21, \infty)$ Por tanto, rechazo H_0, no son homogéneas, sino heterogéneas, la proporción de defectuosos no es la misma para los tres proveedores.

3.4. Otros contrastes no paramétricos

En este apartado veremos algunos procedimientos diseñados especialmente para el contraste de ajuste a distribuciones continuas, aunque los cálculos los realizaremos en prácticas con el R.

El primero de los contrastes considerados es el de Kolmogorov-Smirnov, que se basa en la diferencia máxima entre la función de distribución empírica y teórica: compara la función de distribución teórica F con la empírica F_n mediante el estadístico de contraste:

$$D_n = máx_x |F_n(x) - F(x)|,$$

cuya distribución es independiente del modelo propuesto bajo H_0 y está tabulada. Cuando no se especifican los parámetros, sino que éstos han de estimarse, se debe corregir la distribución del estadístico.

R: ks.test(x, y, ...)

Debido a la gran importancia de la distribución Normal, existen diversos contrastes específicos para estudiar la bondad de ajuste a esta distribución, como:

- Contraste de Shapiro-Wilks: que se basa en el ajuste de la muestra a una recta al dibujarla en papel probabilístico normal. En prácticas, se verá cómo obtener este gráfico (*gráfico de probabilidad normal*).

  R: shapiro.test(x), qqnorm(x), qqline(x)

- Contraste de asimetría, que se basa en que bajo la hipótesis de normalidad, el coeficiente de asimetría poblacional toma el valor cero.
- Contraste de curtosis o apuntamiento, que se basa en que el coeficiente de apuntamiento poblacional de la distribución normal es cero (tras restarle 3).

A continuación se muestra un ejemplo donde se contrasta si unos datos provienen de una distribución normal. Primero, sin especificar a priori sus parámetros.

Ejemplo 3.7 (Examen 1/9/2008): Antes de sacar al mercado un cierto *software* sobre edición de vídeo, se realiza un test de utilización (*usability testing*), en el que potenciales usuarios prueban el producto y se recogen sus datos, para con ellos refinar el producto antes de sacarlo a la venta. Entre los datos recogidos estuvo el tiempo que necesitaron distintos usuarios para completar una determinada tarea de edición, y que se recopiló en el vector Tiempotarea. Usemos los contrastes anteriores para estudiar su posible normalidad ($\alpha = 0.05$).

/**** Contraste de (normalidad) de Shapiro-Wilks ****/

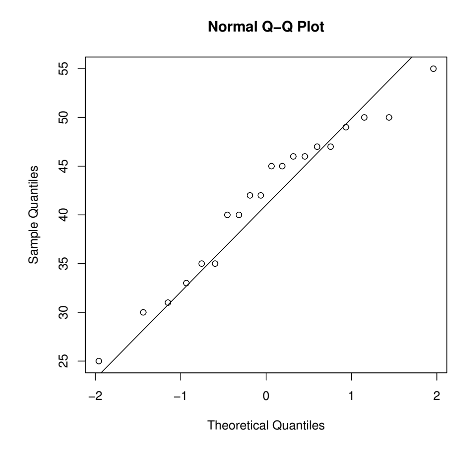
```
> shapiro.test(Tiempotarea)

        Shapiro-Wilk normality test

data:  Disco W = 0.9564, p-value = 0.4747
```

/**** Gráfico de probabilidad normal ****/

```
> qqnorm(Tiempotarea)
> qqline(Tiempotarea)
```



El contraste que planteamos es el siguiente:

$$\begin{cases} H_0: & \text{La variable "Tiempo de completar la tarea" es Normal} \\ H_1: & \text{La variable "Tiempo de completar la tarea" no es Normal} \end{cases}$$

Tanto a través del gráfico (los puntos se ajustan a la recta), como a través del p-valor = 0.4747, mayor que $\alpha = 0.05$, podemos aceptar que los datos de Tiempotarea provengan de una distribución Normal.

Supongamos ahora que deseamos contrastar si es Normal de media 42 y desviación típica 7.

$$\begin{cases} H_0: & \text{La variable "Tiempo de completar la tarea" es Normal(42,7)} \\ H_1: & \text{La variable "Tiempo de completar la tarea" no es Normal(42,7)} \end{cases}$$

/**** Contraste de Kolmogorov-Smirnov ****/

```
> ks.test(Tiempotarea,"pnorm",42,7)

        One-sample Kolmogorov-Smirnov test

 data:  Tiempotarea D = 0.1659, p-value = 0.6409
 alternative hypothesis:two.sided
```

De nuevo, mediante el p-valor = 0.6409, que es bastante elevado ($> \alpha$), podríamos a aceptar que "Tiempo de completar la tarea" siga una distribución Normal(42,7).

Tras aplicar contrastes de normalidad como los anteriores, es posible que no se pueda aceptar que la distribución poblacional sea normal. En ese caso, o bien se utiliza otro modelo paramétrico que se ajuste a los datos o bien se trata de aplicar alguna transformación sobre la variable para tratar de conseguir que los nuevos datos se ajusten a una normal, como son las transformaciones de Box-Cox. Para ello, se pueden consultar [11, 56].

Ejemplo 3.7: También se recogieron los tiempos transcurridos entre llamadas al *help - desk*, en el vector *Tiempohelp*. A continuación aparecen distintas salidas del R, para realizar contrastes de bondad de ajuste a la distribución normal y exponencial (de media 230 minutos), por ese orden. En base a estas salidas, contrasta las hipótesis anteriores.

/**** Normal ****/

```
> shapiro.test(Tiempohelp)

        Shapiro-Wilk normality test

data:  Tiempohelp W = 0.684, p-value = 1.232e-05
```

El contraste es:

$$\begin{cases} H_0: & \text{La variable “Tiempo entre llamadas al help-desk” es Normal} \\ H_1: & \text{La variable “Tiempo entre llamadas al help-desk” no es Normal} \end{cases}$$

El p-valor = 1.232e-05 es muy pequeño ($< \alpha = 0.05$), podemos rechazar claramente que se distribuya normalmente.

/**** Exponencial ****/

```
> ks.test(Tiempohelp,"pexp",1/230)

        One-sample Kolmogorov-Smirnov test

 data:  Tiempohelp D = 0.1873, p-value = 0.4234
 alternative hypothesis: two.sided
```

Fíjate que el parámetro de la exponencial es 1/media.

$$\begin{cases} H_0: & \text{“Tiempo entre llamadas al help-desk” es Exponencial(1/230)} \\ H_1: & \text{“Tiempo entre llamadas al help-desk” no es Exponencial(1/230)} \end{cases}$$

En cambio, ahora el p-valor = 0.4234 es mayor que $\alpha = 0.05$, y podríamos aceptar que los datos procedieran de una Exponencial de media 230 minutos.

Capítulo 4

Control estadístico de calidad

Si Japón puede ... ¿por qué nosotros no podemos? Ellos se dieron cuenta de que los beneficios que tú obtienes mediante los métodos estadísticos, son beneficios que tú obtienes sin nueva maquinaria, sin nuevo personal. Cualquiera puede producir calidad si baja la tasa de producción. Yo no estoy hablando de eso. El pensamiento estadístico y los métodos estadísticos son para los trabajadores, capataces y toda la compañía japonesa, una segunda lengua. Con control estadístico tienes un producto reproducible hora tras hora, día tras día. Y ves qué reconfortante es para la dirección, ellos ahora saben qué pueden producir, ellos saben qué costes habrá.

Los defectos no son gratis. Alguien los hace, y se le paga por hacerlos.

W. Edwards Deming

4.1. Introducción. ¿Qué es el control estadístico de la calidad?

Comencemos por el principio, ¿qué entendemos por calidad? Calidad significa idoneidad de uso. Mejora de la calidad significa la eliminación del desperdicio, lo cual supone mayor productividad, mayor satisfacción del cliente, mayor reputación en la empresa, mayor competitividad y en definitiva, una mayor ganancia.

El control estadístico de la calidad es el conjunto de métodos de ingeniería y estadísticos que se emplean en la medición, vigilancia, control y mejora de la calidad.

¿Por qué hay interés en el control estadístico de la calidad? Son diversas las razones:

1. Incremento en la competitividad entre distintas empresas.

2. Necesidad de evitar pérdidas de material y ahorrar número de horas de las personas.

3. Incremento en el beneficio de la empresa.

4. Incremento en el consumismo. En la decisión de compra de un consumidor, la calidad de un producto puede tener la misma importancia, o superior, que el coste o el tiempo de entrega del mismo.

5. Incremento en las demandas como consecuencia del mal funcionamiento del producto y la necesidad de tener información documentada sobre el proceso de fabricación para una posterior defensa legal frente a demandas de consumidores.

6. La necesidad de conocer la capacidad real del proceso de fabricación.

7. Los cada vez más exigentes requerimientos legales para que el producto pueda ser comercializado.

8. Proliferación de estándares industriales de obligado cumplimiento.

9. Incremento en estándares internacionales para comercio internacional.

Aunque nosotros solamente trataremos el control estadístico de procesos, el control de calidad se clasifica en:

a) Control en curso de fabricación (de procesos).

b) Control de recepción y de producto acabado.

El control en curso de fabricación se realiza durante la fabricación del producto, a intervalos fijos de tiempo, y tiene por objeto vigilar el funcionamiento del sistema y recoger información para mejorarlo. El control de recepción y de producto acabado trata de encontrar una buena manera para decidir si un producto verifica las especificaciones establecidas.

Control de recepción: Un campo donde el muestreo juega un papel fundamental es en el control de recepción que trata de comprobar que los productos cumplan las especificaciones de calidad.

El más empleado es el control de recepción por atributos, en el que se inspeccionan por muestreo las unidades de un lote. Se seleccionan artículos de cada lote y se toma una decisión con base a dicha muestra respecto a si se acepta o se rechaza el lote, según el número de unidades defectuosas que contenga.

Para resolver esta cuestión podemos emplear los llamados planes de muestreo. Éstos podemos clasificarlos en:

a) Planes de aceptación/rechazo: los más conocidos son:

- las normas japonesas JIS Z 9002
- las normas norteamericanas Military Standard MIL- STD- 105D; UNE 66020

Éste último tiene en cuenta el tipo de inspección, así como el rigor de inspección.

Los muestreos pueden ser simples, dobles, múltiples y (en su caso extremo) secuencial (un muestreo es secuencial cuando después de cada observación se toma una de las siguientes decisiones: aceptar el lote, rechazarlo o seguir con el muestreo).

b) Planes de control rectificativo: los lotes rechazados se inspeccionan al 100 % sustituyendo los elementos defectuosos. Los más usados son los de Dodge-Romig.

Las tablas de estos planes y una explicación más detallada sobre muestreo podéis encontrarlos por ejemplo en [56], y sobre todo en cualquier libro sobre Control de Calidad.

4.2. Introducción a los gráficos de control

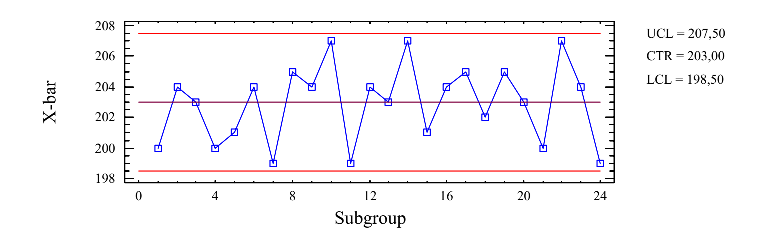
En todo proceso aparece una cierta variabilidad en la calidad, debida a causas aleatorias o no asignables: variabilidad de la materia prima, la precisión de las máquinas y de los instrumentos de medida, destreza de los operarios, etc.

Otras causas no aleatorias o asignables (materias primas defectuosas, desgaste de herramientas, deficiente preparación del operario, etc.) producen ciertos efectos previsibles y definidos. Son pocas y de aparición irregular, pero con grandes efectos. Son eliminables. Diremos que un proceso está en estado de control cuando no le afecta ninguna causa asignable. Un instrumento para determinar si se da o no esta situación son las gráficas de control.

El gráfico de control es una técnica de vigilancia en línea que puede ser utilizada para:

1. La detección rápida de causas asignables.
2. Estimar los parámetros del proceso de producción.
3. Obtención de información para la mejora del proceso, por ejemplo, reduciendo la variabilidad.

Ejemplo 4.1: La empresa el *Girasol azul* se dedica en una de sus plantas al tratamiento de pipas. Éstas se venden en bolsas de 200 gr, a las que se controla su peso. A intervalos de tiempo de 10 minutos, se extraen cuatro bolsas durante la producción y se considera su peso medio, que vamos representando como en la figura siguiente.



Un gráfico de control es una representación de una cantidad (media, rango, proporción, número de defectos, ...) en función del tiempo o número de muestra con unos límites de control. Falta todavía por determinar: ¿qué límites de control son los adecuados? y ¿cuándo un proceso está bajo control?

Si un punto se localiza fuera de los límites de control o aun encontrándose entre los límites de control, si se comportan de manera sistemática o no aleatoria, entonces también es un indicador de que el proceso está fuera de control. Existe una relación entre los gráficos de control y el contraste de hipótesis, estudiado en el tema anterior :

1. En cada punto del gráfico estamos contrastando la hipótesis de que el proceso se encuentra en estado de control estadístico.

2. Tenemos la probabilidad de error tipo I (concluir que el proceso está fuera de control cuando no lo está) y la probabilidad del error tipo II (concluir que el proceso está bajo control cuando no lo está).

Un modelo general de gráfico de control sería como sigue: sea W el estadístico muestral que mide la característica de calidad en la que se tiene interés. Las líneas central (LC), inferior (LIC) y superior (LSC) vienen dadas por

$$\begin{aligned} LSC &= \mu_W + k\sigma_W, \\ LC &= \mu_W, \\ LIC &= \mu_W - k\sigma_W, \end{aligned}$$

donde

$$\begin{aligned} \mu_W &= \text{media de } W, \\ \sigma_W &= \text{desviación típica de } W \end{aligned}$$

y k es la "distancia" de la línea central a los límites de control, siendo $k = 3$ una elección bastante común. Estamos suponiendo que tanto la media, μ_W, como la desviación típica de W son conocidas. Obviamente, habitualmente esto no es así. Los parámetros son típicamente desconocidos y los habremos de estimar a partir de la muestra. La idea de utilizar estos gráficos se debe a Walter A. Shewhart y se habla del *gráfico de control de Shewhart*.

A continuación, clasificaremos los gráficos de control en dos tipos generales.

1. Gráficos de control de variables en donde la característica de calidad es una variable continua. A su vez tendremos gráficos de control para la tendencia central (gráfico $\bar{X}$) y para la variabilidad (gráficos para la desviación típica y para el rango).

2. Gráficos de control de atributos: corresponden a aquellas situaciones en que la característica de calidad no puede ser medida en una escala continua o tan siquiera cuantitativamente. Podemos decidir si la unidad observada es conforme o no sobre la base de verificar o no unos ciertos atributos. O bien, podemos contar el número de defectos que aparece en una unidad de producto.

En otras palabras, lo anterior se resumiría como sigue: el fundamento teórico de una gráfica de control se basa en la construcción, a partir de los valores de la esperanza μ y la desviación típica σ del modelo teórico de distribución que sigue la característica de calidad, de un intervalo de control (generalmente $[\mu$ - 3σ , $\mu + 3\sigma]$). Dentro de este intervalo están casi todos los valores muestrales del proceso, si éste se encuentra bajo control. Las muestras se obtienen a intervalos regulares de tiempo. Un punto que cae fuera de los límites de control, indicaría que el proceso está fuera de control.

El control de calidad se realiza observando en cada elemento:

1) Una característica de calidad medible (longitud, resistencia, contenido de impurezas, etc.) que se compara con un estándar fijado. Es el control por variables (gráficas $\bar{X}$, R, S). La característica se supone distribuida normalmente.

2) Control por atributos:

2.a. Un atributo o característica cualitativa que el producto posee o no (correcto o defectuoso, por ejemplo). Da lugar a las gráficas p y np. La característica se supone distribuida según una Binomial.

2.b. El número total de defectos. Da lugar a las gráficas u y c. La característica se supone distribuida según una Poisson.

Veamos un ejemplo sobre el diseño de una gráfica de control:

Ejemplo 4.2 (Examen 25/1/2006): Una fábrica de papel utiliza gráficos de control para monitorizar diversos aspectos. Supongamos que los papeles deben cortarse según cierta forma y que se controla la longitud entre dos puntos determinados, que llamaremos X, tomando cada vez muestras de tamaño 7. Supongamos que la media y desviación típica del proceso bajo control fuera respectivamente: 2.05 y 0.3, es decir, $X \sim N(\mu = 2.05, \sigma = 0.3)$, y que calculamos la longitud media de cada muestra. La característica que controlamos es:

$$W = \frac{1}{n}\sum_{i=1}^{n} X_i$$

con $n = 7$ de modo que (recuerda el punto 1.7.1, o bien el libro [34])

$$W \sim N(\mu_W = \mu, \sigma_W = \frac{\sigma}{\sqrt{n}}),$$

o sea,

$$W \sim N(2.05, \frac{0.3}{\sqrt{7}}),$$

Los límites de control (k=3) $3 - \sigma$ vienen dados por

$$\begin{aligned} LSC &= \mu_W + k\sigma_W = 2.39 \\ LIC &= \mu_W - k\sigma_W = 1.71 \end{aligned}$$

Problemas básicos con los que nos enfrentaremos será: determinar el tamaño de la muestra y la frecuencia de muestreo.

Lo ideal es mucha muestra tomada con mucha frecuencia, pero en cambio, lo habitual es pequeñas muestras con alta frecuencia. Con el uso cada vez más frecuente de las nuevas tecnologías, los procedimientos automatizados nos irán acercando a una situación en que se muestreará cada ítem.

Sí que debemos poner énfasis en seleccionar los llamados **subgrupos racionales:** seleccionar muestras que, en la medida de lo posible, recojan la variabilidad natural y excluyan la asignable. Obviamente, el orden temporal de la producción será la base lógica para la formación de los subgrupos racionales.

Hay dos opciones básicas para obtener subgrupos racionales:

(i) Cada subgrupo esté formado por unidades producidas al mismo tiempo (o lo más cercanas posibles): daría una instantánea del proceso.

(ii) Cada subgrupo esté formado por unidades que son representativas de todas las unidades producidas desde que se tomó la última muestra.

El primer procedimiento es más sensible a leves corrimientos. Mientras que mediante el segundo podemos decidir si aceptar toda la producción desde última muestra.

Como ya hemos dicho, un gráfico de control puede indicar una condición fuera de control cuando:

(i) Uno o más puntos caen fuera de los límites de control.

(ii) Los puntos exhiben algún patrón no aleatorio de comportamiento.

Se han desarrollado distintos procedimientos empíricos. El más importante sería el de las reglas de la Western Electric. Según estas reglas un proceso está fuera de control cuando:

1. Un punto cae fuera de los límites de control 3-sigma.

2. Dos de tres puntos consecutivos caen fuera de los límites 2-sigma.

3. Cuatro de cinco puntos consecutivos están fuera de los límites 1-sigma.

4. Entendemos que en las dos reglas anteriores los puntos que caen fuera de los límites de control están en el mismo lado, esto es, o son todos mayores que el límite superior correspondiente o menor que el límite inferior.

5. Ocho puntos consecutivos de la gráfica están en el mismo lado de la línea central.

Posteriormente, se volverá a insistir en este punto.

4.3. Gráficos de control para variables

Cuando la característica de calidad es cuantitativa (y asumida normal), controlaremos el valor medio ($\bar{X}$) y la variabilidad (R o S).

Suponiendo que la característica de calidad $X \sim N(\mu, \sigma^2)$, con μ y σ conocidas, entonces la línea central (LC) del gráfico $\bar{X}$ es μ y los límites de control inferior (LIC) y superior (LSC):

$$LSC = \mu + 3\frac{\sigma}{\sqrt{n}},$$

$$LIC = \mu - 3\frac{\sigma}{\sqrt{n}}.$$

¿Qué hacemos cuando no conocemos los parámetros μ y σ?

- Tomamos m muestras previas de tamaño n.
- Si $\bar{X}_i$ es la media de i-ésima muestra entonces μ se estima mediante $\bar{\bar{X}} = \frac{1}{m}\sum_{i=1}^{m} \bar{X}_i$.

Para obtener los límites de control necesitamos la estimación de desviación típica σ: bien utilizando las desviaciones estándar de las distintas muestras o bien utilizando los rangos de dichas muestras. Si, como es habitual n es pequeña (de 4 a 7 son valores usuales), pueden usarse los rangos. Pero, si en lugar de valores pequeños para n, se usan valores mayores que 10 o 12, la estimación de la desviación típica basándonos en el rango es poco eficaz. Empecemos por el caso bastante habitual de usar los rangos.

Supongamos que X_i $i = 1, \ldots, n$ iid normales con μ y σ conocidas. El rango de la muestra aleatoria considerada y el rango relativo serían:

$$R = \max_i X_i - \min_i X_i, \quad W = \frac{R}{\sigma}$$

Los parámetros de la distribución de W dependen sólo del tamaño muestral n, que es conocido:

$$EW = d_2, \quad EW = E\frac{R}{\sigma} = \frac{ER}{\sigma}.$$

La esperanza del rango, ER, la estimamos mediante:

$$\widehat{ER} = \bar{R} = \frac{1}{m}\sum_{i=1}^{m} R_i,$$

donde R_i es el rango correspondiente a la i-ésima muestra. Estimaremos σ mediante (d_2 está tabulada):

$$\hat{\sigma} = \frac{\bar{R}}{d_2}.$$

Por tanto, los límites inferior y superior de control del gráfico $\bar{X}$ son:

$$LSC = \bar{\bar{X}} + \frac{3}{d_2\sqrt{n}}\bar{R}, \;\; LIC = \bar{\bar{X}} - \frac{3}{d_2\sqrt{n}}\bar{R}.$$

Denotando $A_2 = \frac{3}{d_2\sqrt{n}}$ (tabulado), y $\bar{\bar{x}}$, $\bar{r}$ los valores muestrales de los estadísticos:

$$LSC = \bar{\bar{x}} + A_2\bar{r} \;\; LC = \bar{\bar{x}} \;\; LIC = \bar{\bar{x}} - A_2\bar{r}$$

En la tabla siguiente encontramos los valores de los factores que usaremos en este tema:

n	A_2	D_3	D_4	d_2
2	1.880	0	3.267	1.128
3	1.023	0	2.575	1.693
4	0.729	0	2.282	2.059
5	0.577	0	2.115	2.326
6	0.483	0	2.004	2.534
7	0.419	0.076	1.924	2.704
8	0.373	0.136	1.864	2.847
9	0.337	0.184	1.816	2.970
10	0.308	0.223	1.777	3.078

Veamos ahora cómo obtener los límites para el gráfico R.

Si conociésemos la media y la desviación típica del rango, serían:

$$LSC = \mu_R + 3\sigma_R \;\; LC = \mu_R \;\; LIC = \mu_R - 3\sigma_R$$

Si no son conocidos: μ_R será estimada por $\bar{r}$. Como la desviación típica de W, denotada por d_3, es función de n, que es conocida:

$$R = W\sigma \quad \Rightarrow \quad \sigma_R = d_3\sigma.$$

La desviación típica σ es estimada como antes mediante:

$$\hat{\sigma} = \frac{\bar{R}}{d_2},$$

y el estimador de σ_R es:

$$\hat{\sigma}_R = d_3 \frac{\bar{R}}{d_2}.$$

La línea central y límites de control superior e inferior de gráfico R:

$$LSC = \bar{R} + 3\frac{d_3}{d_2}\bar{R} = (1 + 3\frac{\sigma_W}{d_2})\bar{R} \quad LC = \bar{R} \quad LIC = \bar{R} - 3\frac{d_3}{d_2}\bar{R} = (1 - 3\frac{d_3}{d_2})\bar{R}$$

Si denotamos $D_3 = 1 - 3\frac{d_3}{d_2}$ y $D_4 = 1 + 3\frac{d_3}{d_2}$ (tabulados) y sustituimos los estimadores por estimaciones:

$$LSC = D_4\bar{r} \quad LC = \bar{r} \quad LIC = D_3\bar{r}.$$

En resumen:

Los valores del los límites superior e inferior del gráfico de control $\bar{X}$ son:

$$\begin{aligned} LSC &= \bar{\bar{x}} + A_2\bar{r} \\ LC &= \bar{\bar{x}} \\ LIC &= \bar{\bar{x}} - A_2\bar{r} \end{aligned}$$

donde $\bar{\bar{x}} = \frac{1}{m}\sum_{i=1}^{m} \bar{x}_i$ ($\bar{x}_i$ es la media muestral de la muestra i-ésima, calculada con los n valores de cada muestra y m es el número total de muestras), $\bar{r} = \frac{1}{m}\sum_{i=1}^{m} r_i$ (donde r_i es el rango de la muestra i-ésima) y la constante A_2 aparece tabulada.

R: qcc(data, type=“xbar”, center, std.dev, limits, nsigmas = 3, plot = TRUE, ...)

Por otro lado, la línea central y los límites de control superior e inferior de un gráfico R son:

$$\begin{aligned} LSC &= D_4\bar{r} \\ LC &= \bar{r} \\ LIC &= D_3\bar{r}. \end{aligned}$$

Los valores de D_3 y D_4 para distintos valores de n aparecen tabulados.

R: qcc(data, type= “R”, center, std.dev, limits, nsigmas = 3, plot = TRUE, ...)

Un estimador de σ es $\hat{\sigma} = \bar{R}\ /d_2$, donde d_2 está tabulada.

Notemos que estamos estimando unos parámetros asumiendo que las muestras de que disponemos están bajo control. En el caso de que dispongamos de muestras preliminares para llevar a cabo un estudio inicial para la determinación de los límites, procederemos iterativamente como sigue: cálculo de los límites, determinación de las causas asignables si el proceso no ha estado bajo control y reconstrucción del gráfico una vez eliminadas las anomalías y así sucesivamente. Ilustremos la construcción de estos gráficos mediante un ejemplo.

Ejemplo 4.3 (Examen 1/9/2008): Se muestra a continuación las medias y rangos para 15 días, cada uno basado en 6 observaciones diarias del índice de refracción de un cable de fibra óptica.
Medias:
95.7; 95.4; 96.6; 97.4; 96.9; 96.8; 96.5; 98.3; 96; 97.2; 96.5; 96.6; 96.4; 95.5; 97.4
Rangos:
3.2; 6.4; 3.6; 3.2; 1.9; 3.3; 3.4; 3.5; 3.1; 2.3; 3.1; 1.4; 3.8; 1.5; 3.4

1. Utilizando todos los datos, calcula los límites de control para las gráficas de $\bar{X}$ y R.

2. ¿Puede concluirse que el proceso está bajo control? De no ser así, supón que pueden encontrarse las causas asignables, y recalcula los límites de control una vez eliminados los puntos fuera de control.

3. Tras realizar el apartado anterior, estima la media y desviación típica del proceso.

En primer lugar, obtenemos $\bar{\bar{x}} = \frac{1}{m}\sum_{i=1}^{m} \bar{x}_i = \frac{1}{15}\sum_{i=1}^{15} \bar{x}_i = 96.6133$ y $\bar{r} = \frac{1}{15}\sum_{i=1}^{15} r_i = 3.14$. Por otro lado, como $n = 6$, $A_2 = 0.483$, $D_3 = 0$ y $D_4 = 2.004$. En consecuencia, los límites de $\bar{X}$ son:

$$\begin{aligned} LSC &= \bar{\bar{x}} + A_2\bar{r} = 98.131 \\ LC &= \bar{\bar{x}} = 96.6133 \\ LIC &= \bar{\bar{x}} - A_2\bar{r} = 95.0957 \end{aligned}$$

La línea central y los límites de control superior e inferior del gráfico R son:

$$\begin{aligned} LSC &= D_4\bar{r} = 6.29239 \\ LC &= \bar{r} = 3.14 \\ LIC &= D_3\bar{r} = 0. \end{aligned}$$

Vemos que no está bajo control, ya que hay dos puntos fuera de los límites, el 8 en el de la media y el 2 en el del rango. Supongamos que hemos localizado y eliminado la causa asignable que los causó, y éstos puntos vienen eliminados del cálculo de los límites de control, que volvemos a recalcular tras su eliminación (se eliminan de ambas gráficas, ya que si estaban tomados bajo

la presencia de una causa asignable, no estarían representando al proceso bajo control).

En primer lugar, recalculamos $\bar{\bar{x}} = \frac{1}{m}\sum_{i=1}^{m}\bar{x}_i = \frac{1}{13}\sum_{i=1}^{13}\bar{x}_i = (96.6133{\cdot}15 - 98.3 - 95.4)/13 = 96.5769$ y $\bar{r} = \frac{1}{13}\sum_{i=1}^{13} r_i = (3.14 \cdot 15 - 6.4 - 3.5)/13 = 2.86154$. Igual que antes, $n = 6$, $A_2 = 0.483$, $D_3 = 0$ y $D_4 = 2.004$. En consecuencia, los límites de $\bar{X}$ son:

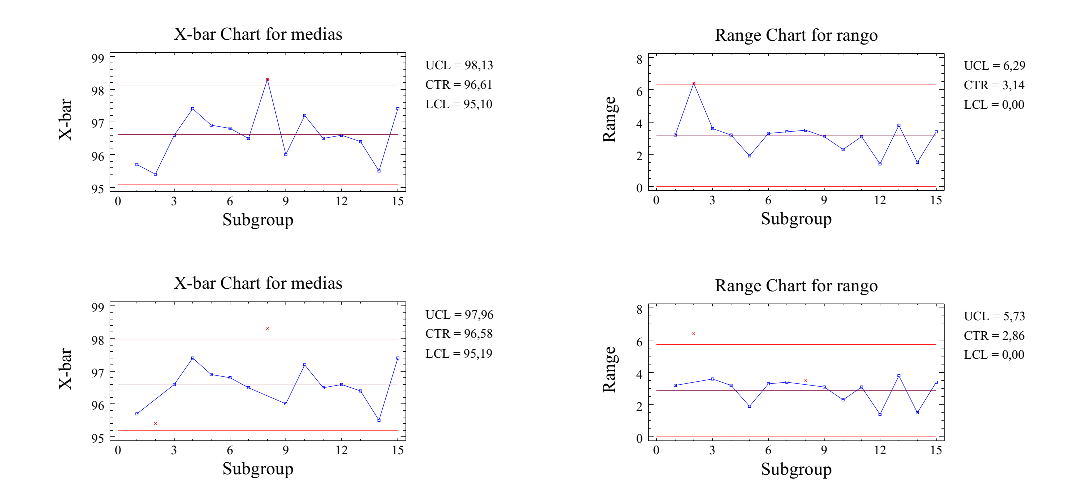
$$\begin{aligned} LSC &= \bar{\bar{x}} + A_2\bar{r} = 97.96 \\ LC &= \bar{\bar{x}} = 96.5769 \\ LIC &= \bar{\bar{x}} - A_2\bar{r} = 95.1939 \end{aligned}$$

La línea central y los límites de control superior e inferior del gráfico R son:

$$\begin{aligned} LSC &= D_4\bar{r} = 5.73437 \\ LC &= \bar{r} = 2.86154 \\ LIC &= D_3\bar{r} = 0. \end{aligned}$$

Ahora ya estaría bajo control y nos quedaríamos con estos límites.

Por último, $\hat{\mu} = 96.5769$, y $\hat{\sigma} = \bar{r}\ /d_2 = 2.86154/2.534 = 1.12926$.



Estos son los gráficos de $\bar{X}$ y R que se obtienen con esos datos, junto con los gráficos recalculados tras eliminar los puntos 2 y 8.

Nota: No confundáis m y n, sobre todo también en caso de tener que recalcular los límites.

Veamos algunas pautas sencillas para interpretar los gráficos de control $\bar{X}$ y R:

a) Puntos fuera de control en $\bar{X}$; R en control: indica un cambio en la media.

b) Puntos fuera de control en $\bar{X}$ y en R: indica un cambio en la variabilidad.

c) Rachas: 7 puntos consecutivos por encima o debajo de la media. Puede indicar (si R está bajo control) cambios en la media (por cambios en la materia prima, el servicio de mantenimiento, etc.).

d) Tendencias: 6 puntos seguidos en sentido creciente o decreciente. Indica la presencia de algún factor que influye gradualmente en el proceso: desgaste de la maquinaria, cambios de temperatura, fatiga (en la gráfica $\bar{X}$); envejecimiento de la maquinaria, mezclas (en R en sentido ascendente); mejora de los operarios o del mantenimiento (en R en sentido descendente).

e) Periodicidades o ciclos: repetición de agrupamientos (sucesión de picos y valles). Indican la presencia de efectos periódicos: temperatura, oscilaciones de corriente (en $\bar{X}$); turnos, acciones de mantenimiento (en R).

f) Inestabilidad: grandes fluctuaciones. Puede indicar un sobreajuste de la máquina, mezcla de materiales, falta de entrenamiento del operario de la máquina.

g) Sobreestabilidad: la variabilidad de las muestras es menor que la esperada (acumulación de puntos en la zona central). Puede que los límites estén mal calculados, que se hayan tomado incorrectamente los datos o que se haya producido un cambio positivo temporal cuya causa debe investigarse.

Como antes hemos dicho, si en lugar de valores pequeños para n, se usan valores mayores que 10 o 12, la estimación de la desviación típica basándonos en el rango es poco eficaz, así que se utilizarán los gráficos $\bar{X}$ y S, cuyos límites pueden calcularse y serían para el gráfico $\bar{X}$:

$$\begin{aligned} LSC &= \bar{\bar{x}} + A_3\bar{S} \\ LC &= \bar{\bar{x}}, \\ LIC &= \bar{\bar{x}} - A_3\bar{S} \end{aligned}$$

con A_3 tabulada y:

$$\bar{S} = \frac{1}{m}\sum_{i=1}^{m} S_i,$$

siendo S_i la desviación típica de la muestra i-ésima. Los límites para el gráfico S serían:

$$LSC = B_4\bar{S},$$
$$LC = \bar{S},$$
$$LIC = B_3\bar{S},$$

nuevamente B_4 y B_3 se encuentran tabuladas.

Veamos su deducción. Ahora, para cada grupo de tamaño n, se calculará la media y la desviación estándar, s:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

S no es un estimador insesgado de σ (S^2 sí estima insesgadamente σ^2):

$$ES = c_4\sigma, \quad c_4 = \sqrt{\frac{2}{n-1}\frac{\Gamma(n/2)}{\Gamma((n-1)/2)}}$$

Además, tenemos que:

$$\sigma_S = \sqrt{var(S)} = \sigma\sqrt{1-c_4^2}.$$

Si conocemos σ (cosa poco frecuente) los límites de control de S son:

$$LSC = c_4\sigma + 3\sigma\sqrt{1-c_4^2} \quad LC = c_4\sigma \quad LIC = c_4\sigma - 3\sigma\sqrt{1-c_4^2}.$$

Denotando, $B_5 = c_4 - 3\sqrt{1-c_4^2}$ y $B_6 = c_4 + 3\sqrt{1-c_4^2}$ (tabulados)

$$LSC = B_6\sigma \quad LC = c_4\sigma \quad LIC = B_5\sigma.$$

Si hay que estimar σ, tendremos m muestras de tamaño n, siendo S_i la desviación típica de la muestra i-ésima y estimaremos ES:

$$\bar{S} = \frac{1}{m}\sum_{i=1}^{m} S_i,$$

y σ lo estimamos mediante:

$$\hat{\sigma} = \frac{\bar{S}}{c_4}.$$

Sustituyendo los valores teóricos por las estimaciones, el gráfico S quedaría:

$$LSC = \bar{s} + 3\frac{\bar{s}}{c_4}\sqrt{1-c_4^2} \quad LC = \bar{s} \quad LIC = \bar{s} - 3\frac{\bar{s}}{c_4}\sqrt{1-c_4^2}.$$

Detonando, $B_3 = 1 - 3\frac{1}{c_4}\sqrt{1-c_4^2}$ y $B_4 = 1 + 3\frac{1}{c_4}\sqrt{1-c_4^2}$, el gráfico de control sería:

$$LSC = B_4\bar{s} \quad LC = \bar{s} \quad LIC = B_3\bar{s}.$$

Notemos que $B_4 = B_6/c_4$ y $B_3 = B_5/c_4$. Como $\hat{\sigma} = \frac{\bar{S}}{c_4}$, el gráfico $\bar{X}$ sería:

$$LSC = \bar{\bar{x}} + 3\frac{\bar{s}}{c_4\sqrt{n}} \quad LC = \bar{\bar{x}} \quad LIC = \bar{\bar{x}} - 3\frac{\bar{s}}{c_4\sqrt{n}}.$$

Denotando $A_3 = \frac{3}{c_4\sqrt{n}}$, el gráfico $\bar{X}$ sería :

$$LSC = \bar{\bar{x}} + A_3\bar{s} \quad LC = \bar{\bar{x}} \quad LIC = \bar{\bar{x}} - A_3\bar{s}.$$

R: qcc(data, type=”S”, center, std.dev, limits, nsigmas = 3, plot = TRUE, ...)

En otras ocasiones, el tamaño muestral es n =1, como en las siguientes:

1. Se utiliza tecnología de medición e inspección automatizada, con lo que se analiza cada unidad producida.
2. El ritmo de producción es lento, y resulta inconveniente permitir que muestras de tamaño $n > 1$ se acumulen antes de ser analizadas.
3. Las mediciones repetidas de un proceso difieren sólo debido a errores en el laboratorio o a errores en el análisis, como sucede en muchos procesos químicos.

Para estimar variabilidad del proceso usamos el rango móvil de dos observaciones consecutivas:

$$MR_i = \mid X_i - X_{i-1} \mid$$

La idea es construir grupos artificiales formados por una observación y la siguiente. Es posible generalizar la idea anterior tomando una observación y tres o cuatro o más consecutivas a ella. Hablaríamos de rangos móviles de orden mayor a uno.

En este caso, la estimación de σ sería:

$$\hat{\sigma} = \overline{MR}/d_2 = (n = 2) = \overline{MR}/1.128$$

El gráfico de control para mediciones individuales será:

$$LSC = \bar{x} + 3\frac{\overline{mr}}{d_2} \qquad LC = \bar{x} \qquad LIC = \bar{x} - 3\frac{\overline{mr}}{d_2}$$

El gráfico de control de rango móvil será:

$$LSC = D_4\overline{mr} = 3.267\overline{mr} \qquad LC = \overline{mr} \qquad LIC = D_3\overline{mr} = 0.$$

R: qcc(data, type="xbar.one", target, center, std.dev, limits, nsigmas = 3, plot = TRUE, ...)

Debido a la dificultad de interpretación del gráfico de rangos móviles por estar correlacionados y la menor sensibilidad del gráfico de mediciones individuales para la detección de pequeños desplazamientos de la media, se verá en este mismo tema, otra alternativa, como es el gráfico de suma acumulada.

El funcionamiento de un proceso en estado de control, no garantiza que sea **capaz** de producir un resultado suficientemente correcto para cumplir los estándares o especificaciones de calidad que se le exijan. No deben confundirse los límites de control, con los límites de las especificaciones, que son externas al proceso.

Vamos a presentar dos índices para comprobar la capacidad de un proceso. En primer lugar, el Índice de la Capacidad del Proceso (ICP):

$$ICP = \frac{LSE - LIE}{6\sigma},$$

donde LSE y LIE son los límites superior e inferior de especificación.

La interpretación del ICP es la siguiente: (1/ICP)100 es el porcentaje del ancho de las especificaciones utilizadas por el proceso, por ello:

(a) $ICP > 1$: pocas unidades defectuosas.

(b) $ICP = 1$: 0.27 % de unidades defectuosas.

(c) $ICP < 1$: muchas unidades defectuosas.

La definición de ICP asume que el proceso está centrado en la dimensión nominal. Es razonable considerar ICP como una medida de la capacidad potencial (si estuviera centrado entonces sí que mediría su capacidad). Por ello, se define ICP_k como un indicador más robusto frente a la no centralidad del proceso:

$$ICP_k = \text{mín}\{\frac{LSE - \mu}{3\sigma}, \frac{\mu - LIE}{3\sigma}\}.$$

En muchas compañías se suele utilizar: ICP = 1.33 de un modo genérico e ICP = 1.66 si la característica de calidad se refiere a resistencia, seguridad u otras características críticas. También se utiliza $ICP_k = 2$ que recibe el nombre de *proceso 6-sigma*, ya que la distancia entre la media y la especificación más cercana es de seis desviaciones estándar.

R: process.capability(object, spec.limits, nsigmas=3, ...)

Ejemplo 4.2: Si las especificaciones son 2 ± 0.1. Calcula los dos índices anteriores, y también la probabilidad de producir unidades por encima, por

debajo de las especificaciones, y en definitiva defectuosas.

LSE = 2 + 0.1 = 2.1 y LIE = 2 - 0.1 = 1.9. Además también sabíamos que $\mu = 2.05$ y $\sigma = 0.3$. Con lo cual:

$$ICP_k = \text{mín}\{\frac{LSE-\mu}{3\sigma}, \frac{\mu - LIE}{3\sigma}\} = \text{mín}\{\frac{2.1-2.05}{3 \cdot 0.3}, \frac{2.05-1.9}{3 \cdot 0.3}\} = 0.055.$$

ICP_k es bajísimo y está muy por debajo de 1.33, es decir, el proceso no es capaz, hay un elevado porcentaje de elementos defectuosos.

Por otro lado,

$$ICP = \frac{LSE - LIE}{6\sigma} = \frac{2.1-1.9}{6 \cdot 0.3} = 0.11$$

La capacidad potencial del proceso, si estuviera (que no lo está) centrado, sería de 0.11 (muy por debajo de 1.33), es decir, el proceso no sería capaz, ni aún estando centrado en la dimensión nominal, y habría también un elevado porcentaje de elementos defectuosos.

Calculemos estos porcentajes. Primero obtendremos la probabilidad de que esté por debajo de las especificaciones, y luego por encima:

$$P(X < LIE) = P(X < 1.9) = P(\frac{X-\mu}{\sigma} < \frac{1.9-2.05}{0.3}) = P(Z < -0.5) = 0.3085$$

$$P(X > LSE) = P(X > 2.1) = P(\frac{X-\mu}{\sigma} > \frac{2.1-2.05}{0.3}) = P(Z > 0.167) = 0.4338$$

En definitiva el porcentaje de defectuosos será: $30.85\,\% + 43.38\,\% = 74.23\,\%$

Una forma de evaluar las decisiones respecto al tamaño de la muestra y la frecuencia de muestreo es a través de la **ARL**, *average run length*, longitud media de la racha. El ARL nos proporcionará el número medio de puntos que deben representarse antes que cualquier punto exceda los límites de control. Si p es la probabilidad de que cualquier punto exceda los límites de control, entonces ARL = $1/p$ (recuerda la distribución geométrica).

Ejemplo 4.2: ¿Cuál es la ARL, para el gráfico $\bar{X}$ con límites 3-sigma?

$$\bar{X} \sim N(2.05, \frac{0.3}{\sqrt{7}}),$$

Los límites de control (k=3) $3-\sigma$ venían dados por

$$\begin{aligned} LSC &= \mu_{\bar{X}} + k\sigma_{\bar{X}} = 2.39 \\ LIC &= \mu_{\bar{X}} - k\sigma_{\bar{X}} = 1.71 \end{aligned}$$

Precisamente por cómo se habían construido los límites:

$$p = P(\bar{X} < LIC) + P(\bar{X} > LSC) = P(Z < -3) + P(Z > 3) =$$

$$2 \cdot 0.00134996 = 0.0027$$

Por tanto, $ARL = 1 \ / \ 0.0027 = 370.37$. Cada 370 muestras o puntos de control tendríamos una falsa alarma, en promedio.

Ejemplo 4.2: Ahora supongamos que el proceso se sale de control y que la media se corre a 1.85. ¿Cuál es la probabilidad de que el desplazamiento se detecte en la primera muestra después del corrimiento? ¿Cuál es la ARL después del corrimiento?

Ahora:

$$\bar{X} \sim N(1.85, \frac{0.3}{\sqrt{7}}),$$

$$p = P(\bar{X} < LIC) + P(\bar{X} > LSC) = P(\bar{X} < 1.71) + P(\bar{X} > 2.39) =$$

$$P(Z < (1.71 - 1.85)/\frac{0.3}{\sqrt{7}}) + P(Z > (2.39 - 1.85)/\frac{0.3}{\sqrt{7}}) =$$

$$P(Z < -1.23) + P(Z > 4.76) = 1 - 0.8906 + 0 = 0.1094$$

ARL= 1 / 0.1094 = 9.14

Si pensamos que es catastrófico darnos cuenta de ese desplazamiento en la novena muestra (en promedio), podemos hacer dos cosas (o una combinación de ambas): muestrear con mayor frecuencia (cada menos tiempo) o bien aumentar el tamaño de las muestras, aumentar n, de esta forma reduciremos la ARL.

4.4. Gráficos de control de atributos

Empecemos con la gráfica P. Es un gráfico de control para la fracción de artículos defectuosos o que no cumplen con las especificaciones, que se basa en la distribución binomial.

Denotamos por p la fracción de piezas no conformes que se producen cuando el proceso está funcionando de un modo estable. Si seleccionamos una muestra de tamaño n y D es el número de unidades no conformes, entonces $D \sim Bi(n, p)$.

Si conocemos p, el gráfico p para la fracción de artículos defectuosos será (recuerda la distribución Binomial):

$$LSC = p + 3\sqrt{\frac{p(1-p)}{n}} \quad LC = p \quad LIC = p - 3\sqrt{\frac{p(1-p)}{n}}$$

Normalmente p es desconocido, así que podemos tomar m (20-25) muestras de tamaño n. Sea D_i el número de unidades no conformes en la muestra i, entonces $\hat{P}_i = \frac{D_i}{n}$ es la proporción de artículos defectuosos en la muestra i.

Estimaremos p mediante $\bar{p} = \frac{1}{m}\sum_{i=1}^{m}\hat{p}_i = \frac{1}{mn}\sum_{i=1}^{m} d_i$.

El gráfico p para la fracción de artículos defectuosos será:

$$LSC = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \quad LC = \bar{p} \quad LIC = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}.$$

Estos límites se han basado en la aproximación normal de la binomial, que podría no ser adecuada si p es pequeño. Si $\hat{p}$ es pequeño, el límite inferior puede ser negativo, en estos casos lo tomaremos como 0.

R: qcc(data, type="p", sizes, center, std.dev, labels, limits, nsigmas = 3, plot = TRUE, ...)

Ejemplo 3.3: Se desea construir un gráfico de control para controlar un proceso que fabrica diodos para un circuito impreso. Se tienen 20 muestras, cada una formada por 50 diodos. El número de diodos defectuosos en cada una de las muestras aparece a continuación: 4 5 3 1 4 5 7 5 10 6 5 1 3 1 2 3 5 4 6 2.

- i) Utilizando todos los datos, calcula los límites de control para un gráfico de control apropiado.
- ii) ¿Puede concluirse que el proceso está bajo control? De no ser así, supón que pueden encontrarse las causas asignables, y recalcula los límites de control una vez eliminados los puntos fuera de control.

En cada muestra, se está monitorizando el número de diodos defectuosos de 50 diodos, es decir, nuestra característica de calidad es una Binomial, por tanto, vamos a usar la gráfica P.

$\bar{p}$ es la estimación de p (fracción defectuosa del proceso), obtenida mediante:

$$\bar{p} = \frac{1}{m}\sum_{i=1}^{m}\hat{p}_i = \frac{1}{20}\sum_{i=1}^{20}\hat{p}_i = 0.082$$

con $\hat{p}_i$ la proporción muestral de unidades defectuosas en la muestra i-ésima (4/50, 5/50, ..., 2/50).

Con lo cual:

$$\begin{aligned} LSC &= \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.082 + 3\sqrt{\frac{0.082(1-0.082)}{50}} = 0.1984 \\ LC &= \bar{p} = 0.082 \\ LIC &= 0.082 - 3\sqrt{\frac{0.082(1-0.082)}{50}} \equiv 0 \end{aligned}$$

Vemos que sólo el punto 9 ($10/50 = 0.2$) estaría fuera de los límites. Lo elimino (suponiendo que se ha detectado y solventado la causa asignable), y recalculo los límites.

$$\bar{p} = \frac{1}{m}\sum_{i=1}^{m}\hat{p}_i = \frac{1}{19}\sum_{i=1}^{19}\hat{p}_i = \frac{82-10}{50\cdot 19} = 0.07579$$

De esta forma:

$$\begin{aligned}
LSC &= \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.07579 + 3\sqrt{\frac{0.07579(1-0.07579)}{50}} = 0.188 \\
LC &= \bar{p} = 0.082 \\
LIC &= 0.07579 - 3\sqrt{\frac{0.07579(1-0.07579)}{50}} \equiv 0
\end{aligned}$$

Ahora, el proceso sí parece estar bajo control, y nos quedaríamos con estos límites.

Cuando un punto muestral caiga fuera de los límites de control, las posibles opciones son:

a) El proceso ha variado, aumentando o disminuyendo (según el sentido del valor extremo) el valor de p.

b) El sistema de medición ha cambiado (el inspector o los criterios de medida).

c) Se ha cometido un error al estimar el valor de p en dicha muestra.

d) El proceso no ha variado, pero los límites de control son erróneos.

e) Nada ha cambiado, simplemente un suceso poco frecuente ha ocurrido.

De igual forma, es posible controlar el número de unidades defectuosas en una muestra. Es incluso más fácil de interpretar por el personal que realiza la inspección. Para ello usaríamos las gráficas NP, con límites (n es el tamaño muestral de cada muestra):

$$\begin{aligned}
LSC &= n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})} \\
LC &= n\bar{p} \\
LIC &= n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}
\end{aligned}$$

R: qcc(data, type="np", sizes, center, std.dev, labels, limits, nsigmas = 3, plot = TRUE, ...)

Si el número de unidades que componen nuestra muestra es variable, podemos contemplar las siguientes dos opciones:

Opción 1: estimamos p globalmente y luego ajustamos la desviación típica de cada muestra considerando su tamaño (límites de control que no son constantes de muestra a muestra). Si denotamos n_i el tamaño de la i-ésima muestra:

$$\bar{p} = \frac{\sum_{i=1}^{m} D_i}{\sum_{i=1}^{m} n_i}.$$

Los límites de control para la muestra i-ésima:

$$LSC = \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}} \quad LC = \bar{p} \quad LIC = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n_i}}.$$

Opción 2: en el caso en que los tamaños no sean muy distintos puede ser más práctico tomar una especie de tamaño promedio que aproxime más o menos bien a todos los puntos, y lo utilizaríamos para todos los puntos:

$$\bar{n} = \frac{\sum_{i=1}^{m} n_i}{m}$$

Cuando lo que interesa es controlar el número de defectos que aparecen en un individuo más que el número de individuos defectuosos, utilizaremos los gráficos C o U, que veremos seguidamente. Por ejemplo, supongamos que revisamos un monitor TFT, concretamente el número de píxeles en mal estado. Si el número de píxeles no es muy grande el producto puede prestar su servicio con una buena calidad. Obviamente, un número excesivo de píxeles que no funcionan adecuadamente será algo desagradable para el usuario y finalmente repercutirá en la venta del mismo.

En el gráfico C controlamos el número total de defectos en una muestra de n unidades, C, asumiéndose que el número de defectos es una distribución de Poisson. Con lo cual, $EC = var(C) = \lambda$.

Los límites (teóricos y desconocidos) del gráfico de control son:

$$LSC = \lambda + 3\sqrt{\lambda} \quad LC = \lambda \quad LIC = \lambda - 3\sqrt{\lambda}$$

Si no conocemos λ tomaremos m muestras, siendo C_i el número de defectos en la i-ésima muestra, y su estimador será: $\hat{\lambda} = \bar{C} = \frac{1}{m}\sum_{i=1}^{m} C_i$.

El gráfico de control C será:

$$LSC = \bar{c} + 3\sqrt{\bar{c}} \quad LC = \bar{c} \quad LIC = \bar{c} - 3\sqrt{\bar{c}}$$

Si el límite de control inferior es negativo tomaríamos el valor 0 en su lugar.

R: qcc(data, type="c", sizes, center, std.dev, labels, limits, nsigmas = 3, plot = TRUE, ...)

Si en cambio queremos controlar los defectos por unidad, usaremos el gráfico U. Utilizamos el promedio de defectos por unidad en la muestra. Si tenemos n (que puede no ser un entero) unidades y un total de defectos C entonces:

$$U = \frac{C}{n},$$

es el promedio de defectos por unidad. Con m muestras preliminares y valores aleatorios $U_1, \ldots, U_m$ entonces el número medio de defectos por unidad es:

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m} U_i.$$

El gráfico de control U es el siguiente (también nos basamos en la aproximación normal de la distribución de Poisson):

$$\begin{aligned} LSC &= \bar{u} + 3\sqrt{\frac{\bar{u}}{n}} \\ LC &= \bar{u} \\ LIC &= \bar{u} - 3\sqrt{\frac{\bar{u}}{n}} \end{aligned}$$

R: qcc(data, type="u", sizes, center, std.dev, labels, limits, nsigmas = 3, plot = TRUE, ...)

Este gráfico, a diferencia del gráfico C, se puede utilizar en aquellos casos en que no se puede tomar una unidad del mismo tamaño para controlar el número de defectos, pudiendo obtener límites no constantes.

Ejemplo 4.2: En la fábrica de papel, se controlan también las imperfecciones en rollos de papel. Se inspeccionan 20 muestras preliminares de 10 rollos cada una, recogiéndose el número de imperfecciones totales: 3 5 7 6 8 9 10 13 6 7 10 9 8 6 5 17 6 15 4 7.

- i) Utilizando todos los datos, calcula los límites de control para una gráfica U.
- ii) ¿Puede concluirse que el proceso está bajo control? De no ser así, supón que pueden encontrarse las causas asignables, y recalcula los límites de control una vez eliminados los puntos fuera de control.

En primer lugar, calculamos:

$$\bar{u} = \frac{1}{m}\sum_{i=1}^{m} u_i = \frac{1}{20}\sum_{i=1}^{20} u_i = \frac{1}{20 \cdot 10}(3 + 5 + 7 + ... + 15 + 4 + 7) = 0.805$$

Así, los límites serían:

$$
\begin{aligned}
LSC &= \bar{u} + 3\sqrt{\frac{\bar{u}}{n}} = 0.805 + 3\sqrt{\frac{0.805}{10}} = 1.656 \\
LC &= \bar{u} = 0.805 \\
LIC &= \bar{u} - 3\sqrt{\frac{\bar{u}}{n}} = 0.805 - 3\sqrt{\frac{0.805}{10}} \equiv 0
\end{aligned}
$$

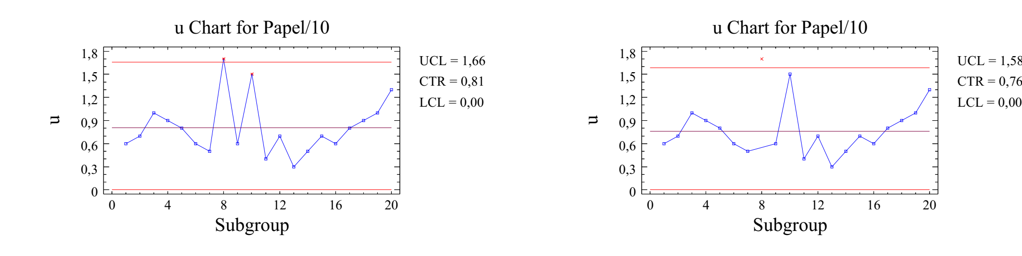
El punto 16 se sale del límite superior ($17/10 = 1.7$), procedemos a su eliminación y recalculamos los límites (suponemos que hemos encontrado las causas asignables).

$$
\bar{u} = \frac{1}{m}\sum_{i=1}^{m} u_i = \frac{1}{19}\sum_{i=1}^{19} u_i = \frac{1}{19 \cdot 10}(161 - 17) = 0.7579
$$

Así los límites serían:

$$
\begin{aligned}
LSC &= \bar{u} + 3\sqrt{\frac{\bar{u}}{n}} = 0.7579 + 3\sqrt{\frac{0.7579}{10}} = 1.584 \\
LC &= \bar{u} = 0.7579 \\
LIC &= \bar{u} - 3\sqrt{\frac{\bar{u}}{n}} = 0.7579 - 3\sqrt{\frac{0.7579}{10}} \equiv 0
\end{aligned}
$$

Estos son los gráficos U que se obtienen con esos datos, junto con el gráfico tras eliminar el punto 16.



Tras la eliminación del 16, el proceso sí parece estar bajo control y nos quedaríamos con estos límites.

4.5. Gráficos de control de suma acumulada

Para finalizar el tema, se presentarán los gráficos de control de suma acumulada (CUSUM), que al igual que los anteriores pueden aplicarse en áreas diversas como control de procesos industriales, administración, ciencias médicas, marketing, comercio, biología, etc.

Estos gráficos surgieron como alternativa a los gráficos Shewhart, que son poco sensibles a pequeños corrimientos de la media (del orden de 1.5σ o inferiores). Esto se debería a que sólo utilizan información del último punto. No consideran toda la secuencia. Alternativas como las reglas de la Western Electric tienen inconvenientes: por un lado, se complica la interpretación del gráfico, y por otro lado, la ARL bajo control se reduce por debajo de 370. Este incremento de las falsas alarmas puede tener consecuencias serias en la práctica.

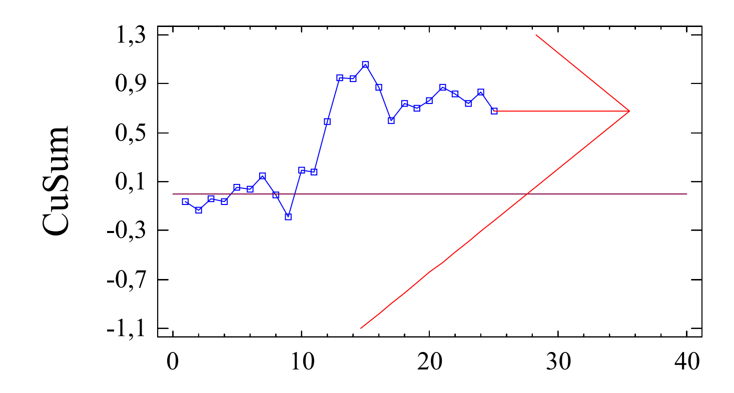
Usaremos los gráficos de la suma acumulada para promedios muestrales y mediciones individuales (para las que son particularmente eficaces), aunque también se pueden plantear para el número de defectos, desviaciones estándar, etc.

Supongamos que μ_0 es el objetivo para la media del proceso y $\bar{X}_j$ es la j-ésima media muestral, entonces el gráfico de control de suma acumulada consiste en representar las sumas dadas por la siguiente ecuación:

$$S_i = \sum_{j=1}^{i} (\bar{X}_j - \mu_0),$$

con $i = 1, \ldots, m$. Notemos que las sumas S_i combinan información de distintas muestras. El punto básico a tener en cuenta es que si el proceso está bajo control alrededor de μ_0, los distintos S_i han de fluctuar alrededor de cero. Si la media se desplaza a μ_1 mayor que μ_0, entonces los S_i tenderán a tomar valores positivos y cada vez mayores. Si la media se desplaza a μ_1 menor que μ_0, entonces los S_i tenderán a tomar valores negativos y cada vez menores. En consecuencia, la observación de una tendencia en el gráfico es un indicativo de que ha habido una modificación de la media y debería buscarse alguna causa asignable.

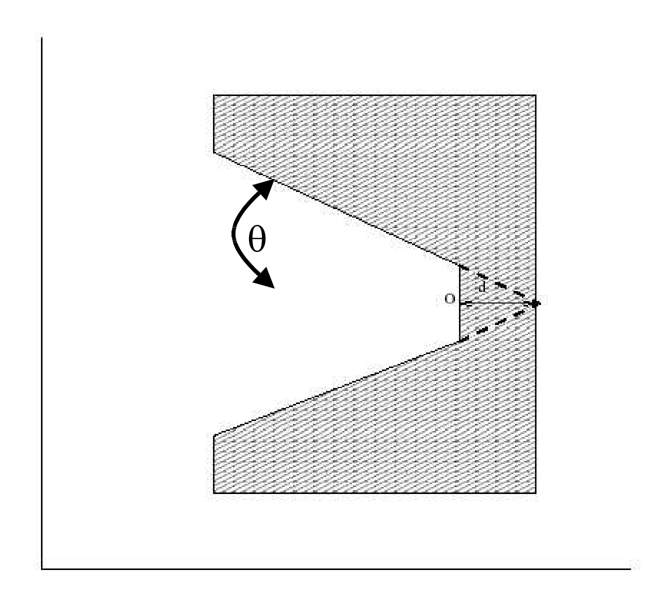
Este gráfico no es una gráfica de control, ya que no tiene límites de control. Dos son los enfoques que se usan para determinar los límites de control: el procedimiento de la máscara V (que vemos en la gráfica siguiente) y el CUSUM tabular, en el que se definen una CUSUM de cola superior e inferior, que acumulan las desviaciones del valor objetivo mayores que cierta cantidad. El proceso estaría fuera de control si exceden cierta constante.



La máscara V viene definida por la distancia d y el ángulo θ, tal y como vemos en la figura siguiente. El origen (O) de la máscara se coloca en la última suma obtenida, y si algún punto queda fuera de los brazos de la V, entendemos que alguna causa asignable ha afectado al proceso. Si α es la probabilidad de error tipo I (falsa alarma), β la probabilidad del error tipo II (no detectamos un corrimiento que sí se ha producido) y Δ el corrimiento mínimo en la media del proceso que deseamos detectar, entonces valores habituales de la máscara V son:

$$d = \frac{2}{\delta^2}\ln(\frac{1-\beta}{\alpha}) \quad \text{y} \quad \theta = \arctan(\frac{\Delta}{2k}) \quad \text{con} \quad \delta = \frac{\Delta}{\sigma_{\bar{X}}}$$

la magnitud del corrimiento expresado en unidades de desviación estándar de la media. k es un factor de escala que relaciona la unidad del eje de ordenadas con la unidad del eje de abscisas (habitualmente k está entre $\sigma_{\bar{X}}$ y $2\sigma_{\bar{X}}$).



Es útil para la programación del procedimiento, su implementación tabular. Tomamos

$$b = \tan(2\theta\sigma_{\bar{X}}) \quad \text{y} \quad h = 2d\sigma_{\bar{X}}\tan\theta$$

Definimos la suma acumulada unilateral superior en el periodo i como:

$$S_H(i) = \text{máx}\{0, \bar{x}_i - (\mu_0 + b) + S_H(i-1)\},$$

y la suma acumulada unilateral inferior como:

$$S_L(i) = \text{máx}\{0, (\mu_0 - b) - \bar{x}_i + S_L(i-1)\}.$$

donde $S_H(0) = S_L(0) = 0$. $S_H(i)$ y $S_L(i)$ acumulan las desviaciones, respecto al valor deseado, que son mayores que b, con ambas cantidades puestas a cero cuando se convierten en negativas. Si $S_H(i)$ o $S_L(i)$ exceden el valor h, entonces el proceso está fuera de control.

En cualquier caso, debemos tener algunas precauciones en la interpretación de los gráficos CUSUM, como son:

1. Controlar la variabilidad (ha de permanecer constante) aparte.
2. No son eficaces en la detección de cambios graduales en la media o que surgen y desaparecen rápidamente. Así que podemos usar CUSUM para detectar saltos en la media y conjuntamente los gráficos Shewhart para ayudarnos a interpretar otras anomalías.

R: object=qcc(data, type="xbar.one", target, ...)
cusum(object)

Capítulo 5

Diseño de experimentos

Llamar al especialista en estadística después de haber hecho el experimento puede no significar más que pedirle que haga un análisis post mórtem: es posible que sea capaz de decir a causa de qué murió el experimento.

Sir Ronald Fisher

Especialmente en África, uno no debe sólo proyectar mejores estadísticas oficiales, sino mejor trabajo experimental en agricultura, medicina e industria.

Gertrude Cox

5.1. Introducción. ¿Qué es el diseño experimental?

En este apartado, se introducen algunos conceptos básicos en experimentación: qué es y cuál es el objetivo del diseño estadístico de experimentos, qué son los factores, los niveles o tratamientos.

Un **experimento** es un conjunto de pruebas o medidas, cuyo objetivo es obtener información que permita tomar decisiones sobre el producto o proceso bajo estudio.

Los experimentos diseñados estadísticamente permiten eficiencia y economía en el proceso experimental, y el empleo de los métodos estadísticos para el análisis de datos, brinda **objetividad científica** a las conclusiones.

Los **factores controlados** en un experimento son las características para las que se prueban diferentes **niveles** o valores con el fin de ver su influencia sobre los resultados. Puede tratarse de factores cuantitativos (temperatura, velocidad, etc.) o cualitativos (proveedor, tipo de máquina, etc.). Los diversos valores de un factor se llaman niveles del factor. En caso de controlar un único factor, sus niveles también se llaman **tratamientos**.

En otras palabras, dentro de un experimento encontraremos: una (aunque podría haber más) variable respuesta (dependiente) y unos factores. Los factores tomarán un número finito de posibles valores, cada uno de ellos se

llamará nivel. Se irán variando los distintos niveles para ver si influyen sobre la variable respuesta. En definitiva, los modelos de *diseño de experimentos*, estudian la variabilidad de la variable de interés controlando los factores que pueden influir en la misma, frente a los modelos de *regresión*, que estudian la variabilidad de la variable de interés teniendo en cuenta la relación funcional de ésta con otras variables explicativas (que normalmente son continuas y en muchos casos no controlables). Aunque estos dos tipos de modelos lineales pueden estudiarse con una visión unificada.

Ejemplo 5.1: Queremos probar 3 tipos de ventiladores de 3 proveedores, para comprar el que menos ruido produzca (otra característica que podría ser de interés, sería el que más enfríe, es decir, también podría interesarnos la temperatura). Experimento: seleccionamos 10 ventiladores de cada proveedor (factor). Colocamos los 30 ventiladores en el mismo ordenador aleatoriamente, ejecutando 1 hora el mismo programa y se mide el sonido (variable respuesta).

A grandes rasgos, en un diseño experimental encontraremos las siguientes etapas:

a) Definir la característica (o características) sobre las que se quiere investigar los posibles efectos de los factores (las respuestas).

b) Seleccionar los factores a incluir en el experimento.

c) Seleccionar los niveles (cuántos y cuáles) considerar en cada factor.

d) Definir en qué va a consistir cada prueba.

e) Decidir el número de pruebas a realizar y el tratamiento a aplicar en cada una (elección del diseño experimental).

f) Organizar todo el trabajo experimental, asignando las responsabilidades correspondientes y precisando las necesidades de tiempo y recursos.

g) Analizar estadísticamente los resultados, para obtener respuesta (y sacar las conclusiones pertinentes) a preguntas como las siguientes: ¿qué factores tienen un efecto significativo sobre la media de la respuesta?, ¿qué interacciones son significativas? (interacción: implica que el efecto de un factor sobre la respuesta es diferente según el nivel al que se halle otro factor considerado), ¿cuáles serían los niveles óptimos para los diferentes factores, en función de sus efectos sobre la media de la respuesta?, ¿qué respuesta media cabe predecir trabajando en las condiciones óptimas encontradas?, ¿hay efectos significativos sobre la varianza de la respuesta?, ¿debemos ampliar el experimento, si hubiera algún punto oscuro?

En resumen, un problema de diseño experimental comprenderá los siguientes puntos: comprensión y planteamiento del problema, elección de factores y

niveles, selección de la variable respuesta, elección del diseño experimental, realización del experimento, análisis de los datos, conclusiones y recomendaciones.

Además, el diseño experimental utiliza tres principios básicos: obtención de réplicas, aleatorización y análisis por bloques. Veamos por orden el significado de estos tres principios, usando el siguiente ejemplo.

Ejemplo 5.2: Se trata de comparar la eficiencia (tiempos de ejecución) de tres algoritmos en la resolución de un cierto problema numérico.

Repetición: Se asignan los mismos tratamientos a las diferentes unidades experimentales (elementos de un diseño). **Ejemplo 5.2:** obviamente, si sólo tuviéramos una medida para cada algoritmo, no tendríamos manera de establecer la variabilidad natural y el error de medida. Sin la repetición, sería imposible.

Aleatorización: ¡Es el paso fundamental de todas las estadísticas! Las unidades experimentales se asignan al azar a los tratamientos o niveles. **Ejemplo 5.2:** si tomáramos primero todas las medidas para el primer algoritmo, luego para el segundo y por último, para el tercero, en vez de aleatoriamente, nos podría ocurrir que algún factor no controlado, como calentamiento de la CPU o procesos que no controlamos, nos afectaran a los resultados. La aleatorización es nuestra manera de *cancelar* los factores no controlados.

Control local: Hace referencia a cualquier método que represente y reduzca la variabilidad natural. Una de sus formas es la agrupación de las unidades experimentales en bloques (cuando tratamos muestras apareadas ya estábamos utilizando este principio). **Ejemplo 5.2:** como los tiempos de ejecución se ven afectados por la elección del *hardware*, cada algoritmo debería ejecutarse en distintas máquinas, cada una de las cuales sería un bloque.

5.2. Diseño completamente aleatorizado: análisis de la varianza con un solo factor

En este apartado, veremos los diseños con una fuente de variación. El problema más sencillo que puede presentarse es el de detectar la influencia de un factor que tiene dos niveles, en una variable de interés. Éste sería el mismo problema que el de comparar las medias de dos poblaciones, que bajo la hipótesis de normalidad podemos resolver mediante el contraste de la t, como se ha visto en el tema sobre Contraste de hipótesis. La generalización de este problema, es contrastar la igualdad de las medias de los a niveles de un factor, es decir, estudiar la influencia de un factor con a niveles en la variable de interés. Veamos un ejemplo de este diseño.

Ejemplo 5.3 (Examen 21/1/2009): Imagina que disponemos de un programa para simular el sistema y obtener los resultados (asumidos normales), en el contexto siguiente: control de la producción. En la tabla siguiente aparecen los costes (gastos) totales medios por mes para cinco réplicas independientes de 4 políticas de inventario diferentes:

Política	Observaciones				
Política 1: (20,40)	126.9	124.3	126.7	122.6	127.3
Política 2: (20,80)	118.2	120.2	122.4	122.7	119.4
Política 3: (40,60)	120.7	129.3	120.6	123.6	127.3
Política 4: (40,100)	131.6	137	129.9	129.9	131

En un diseño experimental completamente aleatorizado, describiremos las observaciones con el modelo estadístico lineal (se considera que el factor de interés tiene a niveles y que inicialmente el número de observaciones, n, es igual para cada tratamiento):

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad i = 1, \ldots, a \quad j = 1, ..., n,$$

donde Y_{ij} es una variable aleatoria que denota la observación ij-ésima, μ_i sería la media del tratamiento i y para los errores $\{\epsilon_{ij}\}$ haremos las siguientes suposiciones:

- Tienen esperanza nula.
- Su varianza es siempre constante, σ^2.
- Tienen una distribución normal.
- Son independientes entre sí.

Una formulación alternativa de estas hipótesis es:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2),$$

con τ_i definida como desviaciones de la media global μ, por lo que $\sum_{i=1}^{a} \tau_i = 0$.

La hipótesis de interés en este tipo de problemas es la de que *no hay diferencias significativas entre los niveles del factor*, que queda formalmente expresada por $H_0 : \mu_1 = \ldots = \mu_a$ (y H_1: no todas las medias son iguales) o equivalentemente $H_0 : \tau_i = 0 \ \forall i$.

La idea (que intuitivamente puede verse en el gráfico siguiente: en el primer caso no está claro si hay diferencia entre los dos grupos, en cambio en el segundo caso, es claro que sí las hay, mientras que en el último caso no habría diferencia) para deducir el estadístico de contraste se basa en descomponer la *variabilidad total* de los datos en dos términos: la *variabilidad entre* las medias de cada muestra y la media general, y la *variabilidad dentro* de cada grupo o *residual*:

$$\begin{array}{ccccc} \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y})^2 & = & n \sum_{i=1}^{a} (\bar{y}_i - \bar{y})^2 & + & \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2 \\ SC_T & = & SC_{Tratamientos} & + & SC_E \end{array}$$

En esta identidad, y_{ij} representa la observación j-ésima obtenida en el tratamiento i-ésimo, $\bar{y}_{..}$ es la media de todas las observaciones, $\bar{y}_{i.}$ es la media de las observaciones bajo el tratamiento i-ésimo. Además, SC_T denota la suma de cuadrados total, $SC_{Tratamientos}$ la suma de cuadrados de los tratamientos (*variabilidad entre*) y SC_E la suma de cuadrados del error (*variabilidad dentro*).

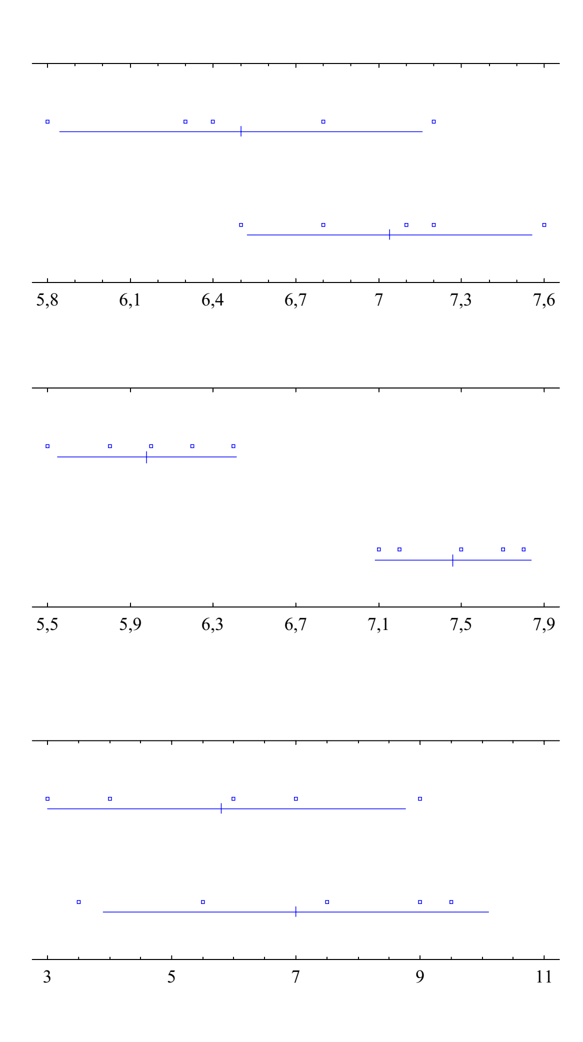


Figura 5.1: Idea intuitiva de ANOVA

La $SC_{Tratamientos}$ mediría la variabilidad explicada por las diferencias entre las medias de tratamientos, mientras que SC_E mediría la variabilidad no explicada. Cuando haya diferencias reales entre las medias en cada nivel, la variabilidad *entre* será grande, comparada con la variabilidad residual. Juzgar su tamaño relativo requiere conocer su distribución en el muestreo.

Se demuestra que cuando H_0 es cierta, SC_E/σ^2 y $SC_{Tratamientos}/\sigma^2$ son variables independientes que tienen una distribución ji-cuadrado con $a(n-1)$ y $a-1$ grados de libertad respectivamente.

El estadístico de contraste será pues:

$$F_0 = \frac{SC_{Tratamientos}/(a-1)}{SC_E/a(n-1)}$$

que tendrá una distribución F de Snedecor con $a-1$ y $a(n-1)$ grados de libertad. El numerador es conocido como cuadrado medio de los tratamientos ($CM_{Tratamientos}$) y el denominador como cuadrado medio del error (CM_E). Puede demostrarse también, que CM_E es un estimador insesgado de σ^2. Por otra parte, si la hipótesis nula es cierta, $CM_{Tratamientos}$ es un estimador insesgado de σ^2. Sin embargo, si la hipótesis nula es falsa, el valor esperado de $CM_{Tratamientos}$ es mayor que σ^2. Por consiguiente, H_0 se rechazará al nivel α si $f_0 > F_{\alpha,a-1,a(n-1)}$, es decir, tendremos un contraste unilateral en el que la región crítica es la cola derecha de la distribución F.

Los términos de la descomposición en que se basa este contraste suelen disponerse en una tabla conocida como *tabla ANOVA* (según la arraigada terminología anglosajona: $ANalysisOfVAriance$). En la tabla 5.1 podemos ver la tabla ANOVA para este primer modelo. En esta tabla aparece el análisis de varianza cuando contamos con un diseño desbalanceado o desequilibrado (el número de observaciones en cada tratamiento puede ser diferente), sólo deben realizarse ligeras modificaciones en las fórmulas anteriores de las sumas de cuadrados. Elegir un diseño balanceado tiene dos ventajas: 1) si los tamaños son iguales, el procedimiento es relativamente insensible a las pequeñas desviaciones del supuesto de la igualdad de varianzas y 2) la potencia de la prueba se maximiza si las muestras tienen igual tamaño. Denotaremos por n_i las observaciones en el tratamiento i-ésimo y N el total de observaciones.

Fuente de variación	Suma de cuadrados	Grados de libertad	Media de cuadrados	F
Tratamientos (entre grupos)	$\sum_{i=1}^{a} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	$a-1$	$\frac{SC_{Trat.}}{(a-1)}$	$\frac{CM_{Trat.}}{CM_E}$
Error (dentro grupos)	$\sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$	$N-a$	$\frac{SC_E}{N-a}$	
Total	$\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$	$N-1$		

Tabla 5.1: Tabla ANOVA de un factor

Región crítica (a nivel α): $(F_{\alpha,a-1,N-a}, \infty)$

R: a = aov(respuesta $\sim$ factor)
anova(a)

Ejemplo 5.3: Especifica el modelo y prueba la igualdad de los efectos (usa $\alpha = 0.05$), explicando el resultado que obtengas.

Se trata de análisis de la varianza con un factor (Política), siendo la variable respuesta Costes:

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad i = 1, \ldots, 4 \quad j = 1, ..., 5,\ \epsilon_{ij} \sim N(0, \sigma^2)$$

Para resolver el contraste:

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_1: & \text{No todas las medias son iguales} \end{cases}$$

construimos la tabla ANOVA.

Supongamos que tenemos parte de la tabla, y tenemos que acabar de rellenar los huecos:

```
--------------------------------------------------------------
Fuente      Suma Cuadrados    GL      Cuadrados medios     F
--------------------------------------------------------------
Política

Error                                     7,9705
--------------------------------------------------------------
TOTAL          459,172
--------------------------------------------------------------
```

Podemos obtener los grados de libertad fácilmente. Serían respectivamente, a - 1 = 4 - 1 = 3, $N - a$ = 20 - 4 = 16, N - 1 = 20 - 1 = 19. A partir de ahí, podemos obtener $SC_E = 16 \cdot 7.9705 = 127.53$. Con lo cual podemos también deducir el valor de $SC_{Tratamientos}$ = 459.172 - 127.53 = 331.64. Entonces, podemos obtener $CM_{Tratamientos}$ = 331.64 / 3 = 110.55, y por último, F = 110.55 / 7.97 = 13.87.

En definitiva, la tabla ANOVA (según la salida del R) quedaría:

```
Analysis of Variance Table

Response: Costes
                  Df Sum Sq Mean Sq F value    Pr(>F)
factor(Politica)  3 331.64  110.55  13.870 0.0001026 ***
 Residuals       16 127.53    7.97
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Región crítica (a nivel α): $(F_{\alpha,a-1,N-a}, \infty) = (F_{0.05,3,16}, \infty) = (3.24, \infty)$

$13.87 \in (3.24, \infty)$ con lo cual, rechazo H_0, sí que hay diferencia entre los costes medios (μ_i), según la política de inventario usada.

Para estimar los parámetros del modelo, que usaremos para construir los residuos y comprobar la validez del modelo, podemos hacerlo mediante:

$\hat{\mu} = \bar{y}_{..} = \frac{\sum_i \sum_j y_{ij}}{N}$

$\hat{\mu}_i = \bar{y}_{i.} = \frac{\sum_j y_{ij}}{n_i}$

$\hat{\tau}_i = \hat{\mu}_i - \hat{\mu}$, donde se usa $\sum_{i=1}^{a} n_i \tau_i = 0$.

Los residuos ($e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i.}$) son útiles para verificar las hipótesis básicas del modelo: comprobar su normalidad (recuerda los contrastes vistos en el capítulo 3), comprobar la variabilidad constante, representando los residuos frente a los valores ajustados o frente a los niveles del factor, la homogeneidad de varianzas también puede comprobarse mediante algún test como el de Bartlett, representar los residuos frente al tiempo (por si hubiera alguna traza de no independencia), comprobar si existen valores atípicos. En [55, pág. 58] se resumen los efectos de las desviaciones en las hipótesis básicas.

En la figura 5.2 se muestran gráficos de residuos frente a valores previstos (media prevista por el modelo para dicho punto). En ambas se incumplen las hipótesis del modelo, en el primero la variabilidad va aumentando con el nivel y en el segundo, la variabilidad de un grupo es mucho mayor que en los restantes.

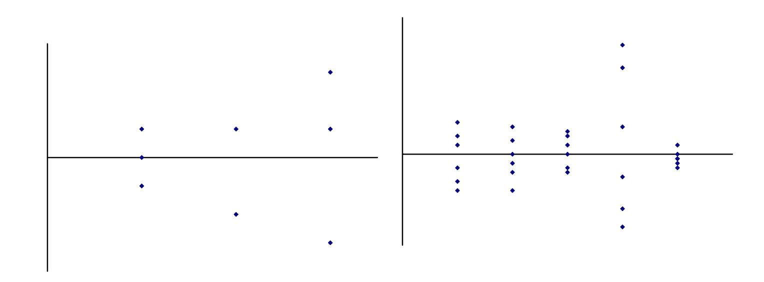


Figura 5.2: Diagramas de residuos frente a valores previstos

Por otro lado, en la figura 5.3, se representan los residuos en su secuencia temporal de obtención, mostrando cuatro situaciones interesantes. En la primera se ha producido a partir de cierto instante un aumento en la respuesta, en el gráfico contiguo se observa una correlación negativa entre los residuos, en el primer gráfico inferior el cambio es gradual y aparece una tendencia, mientras que en el último gráfico se muestra una reducción paulatina de la variabilidad (efecto de aprendizaje).

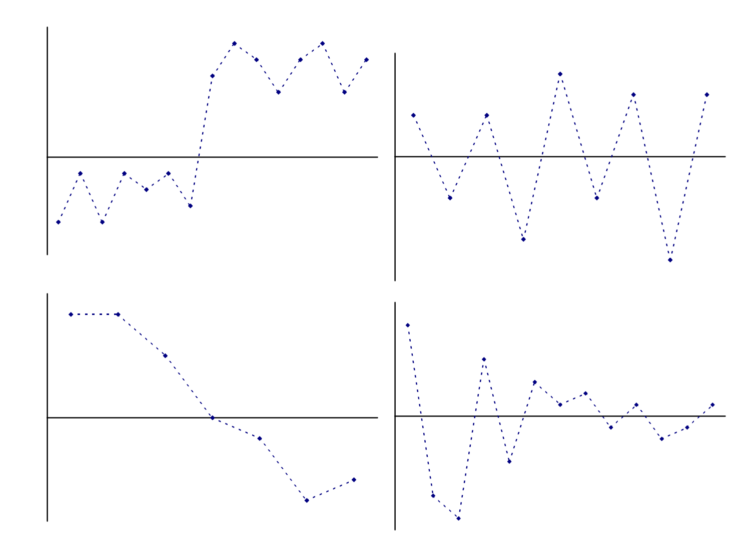


Figura 5.3: Diagramas de residuos en función del tiempo

Seguidamente, se verán algunos métodos para comparar las medias, cuando el efecto ha sido declarado significativo por la prueba F. Podrían dividirse en comparaciones a priori y a posteriori.

En las *comparaciones a priori* (nosotros no las veremos en este curso), antes de llevar a cabo el experimento, ya se saben las comparaciones de interés: se considerarán los contrastes ortogonales. Un contraste lo podemos plantear como $H_0 : \sum_i c_i\mu_i = 0$ con $c_1, \ldots, c_a$ constantes conocidas con $\sum_i n_i c_i = 0$. Dos contrastes con coeficientes $\{c_i\}$ y $\{d_i\}$ serán ortogonales si $\sum_i n_i c_i d_i = 0$. Para probar un contraste se debe comparar su suma de cuadrados $SC_C = (\sum_{i=1}^{a} c_i y_{i.})^2 / \sum_{i=1}^{a} n_i c_i$ ($y_{i.}$ es el total de las observaciones del tratamiento i-ésimo) con la media de cuadrados del error. El estadístico resultante tiene una distribución F con 1 y $N - a$ grados de libertad.

Por otro lado, si no se tiene planteada ninguna pregunta con respecto a las medias de los tratamientos (*comparaciones a posteriori*), se presentará el método de la mínima diferencia significativa o LSD (*Least Significant Difference*) (otros métodos pueden encontrarse en [47]). De esta manera, se compararán todos los pares de medias con las hipótesis nulas $H_0 : \mu_i = \mu_j$ (para toda $i \neq j$) y el par de medias μ_i y μ_j se declarará significativamente diferente si $|\bar{y}_{i.} - \bar{y}_{j.}| >$ LSD, donde LSD al nivel α viene definida como $t_{\alpha/2,N-a}\sqrt{CM_E(1/n_i + 1/n_j)}$. Fíjate que si tenemos el mismo número de observaciones para el tratamiento i-ésimo y j-ésimo, n_i será igual a n_j.

Se pueden representar grupos homogéneos con columnas de X's, de forma que si están en la misma columna no existirían diferencias estadísticamente significativas entre las medias.

R: multicomp.lm(objeto aov, method=“lsd”,error.type=“cwe”)

Ejemplo 5.3: Calcula la LSD al 95 % y realiza las comparaciones explicando los resultados. Identifica grupos homogéneos con columnas de X's. ¿Cuál/es política/s serían mejores?

En este ejemplo, como todas las políticas tienen el mismo número de observaciones, la LSD será la misma para todos los pares de medias:

$\text{LSD} = t_{\alpha/2,N-a}\sqrt{CM_E(1/n_i + 1/n_j)} = t_{0.05/2,16}\sqrt{7.97(1/5+1/5)} =$
$2.12 \cdot \sqrt{7.97(1/5+1/5)} = 3.785$

Las medias observadas para cada tratamiento (política) son:

$\bar{y}_{1.} = 125.56$, $\bar{y}_{2.} = 120.58$, $\bar{y}_{3.} = 124.3$, $\bar{y}_{4.} = 131.88$

Realizamos las comparaciones por parejas:

1-2: $|\bar{y}_{1.} - \bar{y}_{2.}| = 4.98 > \text{LSD} \rightarrow \mu_1 \neq \mu_2$
1-3: $|\bar{y}_{1.} - \bar{y}_{3.}| = 1.26 < \text{LSD} \rightarrow \mu_1 = \mu_3$
1-4: $|\bar{y}_{1.} - \bar{y}_{4.}| = 6.32 > \text{LSD} \rightarrow \mu_1 \neq \mu_4$
2-3: $|\bar{y}_{2.} - \bar{y}_{3.}| = 3.72 < \text{LSD} \rightarrow \mu_2 = \mu_3$
2-4: $|\bar{y}_{2.} - \bar{y}_{4.}| = 11.3 > \text{LSD} \rightarrow \mu_2 \neq \mu_4$
3-4: $|\bar{y}_{3.} - \bar{y}_{4.}| = 7.58 > \text{LSD} \rightarrow \mu_3 \neq \mu_4$

Ordenamos las políticas, según sus medias observadas, de menor a mayor, y colocamos una X en distinta columna si existe diferencia entre las medias poblacionales (fíjate que como la 3 no difiere de la 2 ni de la 1, pero la 1 y la 2 si son distintas, la 3 tiene dos X's para poder expresarlo):

```
Política                  Media
2                         120,58          X
3                         124,3           XX
1                         125,56           X
4                         131,88            X
```

Las mejores políticas serían la 2 y la 3, que forman un grupo homogéneo. A continuación, vendría el grupo formado por la 3 y la 1. Mientras que en el último lugar, estaría la política 4, que sería la peor (mayor gasto medio).

Una alternativa no paramétrica de la prueba F, es el test de Kruskal-Wallis. Únicamente se requiere que las ϵ_{ij} tengan la misma distribución continua para todos los niveles del factor. Esta prueba se basa en los rangos (orden) de las observaciones y el estadístico de la prueba es:

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{a} n_i(\bar{R}_{i.} - \frac{N+1}{2})^2 = \frac{12}{N(N+1)}\sum_{i=1}^{a}\frac{R_{i.}^2}{n_i} - 3(N+1),$$

donde $R_{i.}$ es el total de los rangos del tratamiento i-ésimo y $\bar{R}_{i.}$ la media. Se rechazará H_0 si el valor observado $h \geq \chi^2_{\alpha,a-1}$, con un nivel de significación aproximado α.

5.3. Diseño en bloques aleatorizados

El siguiente modelo que se presenta es el diseño en bloques aleatorizados. En el anterior modelo los factores no controlados por el experimentador y que podían influir en los resultados se asignaban al azar a las observaciones. En este modelo, las unidades experimentales han sido agrupadas según otra causa de variabilidad que puede influir en los resultados: es una variable, denominada *variable de bloqueo*, cuyo efecto sobre la respuesta no es directamente de interés, pero de esta manera se obtienen comparaciones homogéneas, de forma análoga al procedimiento de la prueba t apareada. En este diseño, tomaremos el mismo número de muestras por tratamiento dentro de cada bloque y el orden de las medidas dentro del bloque se decidirá aleatoriamente.

En otras palabras, en este diseño se considera un factor para ver su influencia sobre la variable respuesta y una variable de bloqueo, llamada bloque, que hará las comparaciones más homogéneas. En realidad, no se está interesado en la variable bloque, pero despreciar su influencia (no considerar esta variable) podría ser perjudicial, análogamente a como ocurría con las muestras apareadas en los temas anteriores. El factor bloque es una variable que suponemos que influye en la respuesta, aunque no estamos interesados en conocer su influencia.

Las hipótesis básicas del modelo (sin repetición) son:

$$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \quad i = 1, \ldots, a \quad j = 1, \ldots b,$$

donde ϵ_{ij} son variables $N(0, \sigma^2)$ independientes.

El modelo descompone la respuesta en:

- Una media global μ.
- El efecto incremental en la media debida al nivel del factor, τ_i ($\sum_i \tau_i = 0$).
- El efecto incremental en la media debida al bloque, β_j ($\sum_j \beta_j = 0$).
- El error experimental, ϵ_{ij}, que recoge el efecto de todas las restantes causas posibles de variabilidad del experimento.

Notar que este modelo supone que los efectos del factor y de la variable de bloqueo son aditivos, es decir, no existe interacción entre ambos.

Al igual que antes, resulta interesante ilustrar este diseño con un ejemplo.

Ejemplo 3.7: En otro punto del estudio, se quería comparar 3 distintas interfaces. Para ello, a 5 sujetos se les midió el tiempo empleado en completar

una determinada tarea para cada una de las interfaces, cuyo orden de presentación fue aleatoriamente seleccionado. A continuación, aparecen estos datos:

	Sujetos				
Interfaz	Suj. 1	Suj. 2	Suj. 3	Suj. 4	Suj. 5
1	55	49	43	36	45
2	60	53	41	40	55
3	51	46	39	35	45

La hipótesis de interés será $H_0 : \tau_i = 0 \ \forall i$, pero en algunas ocasiones también puede ser de interés contrastar $H_0 : \beta_j = 0 \ \forall j$.

La deducción de los estadísticos de contraste y sus distribuciones muestrales se obtiene de una manera completamente análoga al anterior. Así, la *variabilidad total* puede descomponerse en *variabilidad entre tratamientos*, *variabilidad entre bloques* y el *error* o *variabilidad dentro de los tratamientos y bloques*.

$$\begin{array}{cc} \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{..})^2 & = \\ SC_T & = \end{array}$$

$$\begin{array}{ccccc} b\sum_{i=1}^{a}(\bar{y}_{i}-\bar{y})^2 & +a\sum_{j=1}^{b}(\bar{y}_{.j}-\bar{y})^2 & +\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij}-\bar{y}_{.j}-\bar{y}_{i.}+\bar{y})^2 \\ SC_{Tratamientos} \quad + & SS_{Bloques} \quad + & SC_E \end{array}$$

Las fórmulas para el caso de un diseño por bloques aleatorizados con repetición pueden consultarse en [23], pero no se tratarán.

Puede demostrarse que: $\mathrm{E}(CM_{Tratamientos}) = \mathrm{E}(SC_{Tratamientos}/a-1) = \sigma^2 + b\sum_{i=1}^{a}\tau_i^2/a-1$, $\mathrm{E}(CM_{Bloques}) = \mathrm{E}(SC_{Bloques}/b-1) = \sigma^2 + a\sum_{j=1}^{b}\beta_j^2/b-1$ y $\mathrm{E}(CM_E) = \mathrm{E}(SC_E/(a-1)(b-1)) = \sigma^2$. Por tanto, la hipótesis nula de que todos los efectos de los tratamientos son cero, se rechazará con el nivel de significación α, si el valor calculado del estadístico $F_0 = CM_{Tratamientos}/CM_E > F_{\alpha,a-1,(a-1)(b-1)}$. En la tabla 5.2 podemos ver la tabla de ANOVA para este modelo.

Fuente de variación	Suma de cuadrados	Grados de libertad	Media de cuadrados	F
Tratamientos	$SC_{Tratamientos}$	$a-1$	$\frac{SC_{Tratamientos}}{(a-1)}$	$\frac{CM_{Tratamientos}}{CM_E}$
Bloques	$SC_{Bloques}$	$b-1$	$\frac{SC_{Bloques}}{b-1}$	
Error	SC_E	$(a-1)(b-1)$	$\frac{SC_E}{(a-1)(b-1)}$	
Total	SC_T	$ab-1$		

Tabla 5.2: Tabla ANOVA de un diseño en bloques aleatorizados

Región crítica (a nivel α): $(F_{\alpha,a-1,(a-1)(b-1)}, \infty)$

R: a = aov(respuesta ~ factor1 + factor bloques)
anova(a)

Ejemplo 3.7: Especifica el modelo y prueba la igualdad de los efectos ($\alpha = 0.05$).

Se trataría de un diseño en bloques aleatorizados, donde el factor de interés es Interfaz, los bloques son los sujetos y la respuesta, el Tiempo empleado en realizar la tarea. Si asumimos que se cumplen las hipótesis del modelo (ϵ_{ij} son variables $N(0, \sigma^2)$ independientes), la respuesta la descompondríamos en:

$$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \quad i = 1, \ldots, 3 \quad j = 1, \ldots 5,$$

Si denotamos por $\mu_i = \mu + \tau_i$, entonces el contraste sería:

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \mu_3 \\ H_1: & \text{No todas las medias son iguales} \end{cases}$$

Construimos la tabla ANOVA siguiente, pues se quiere estudiar si hay diferencias significativas entre las interfaces.

Fuente	Suma de cuadrados	GL	Cuadrados medios	F
Interfaz	111,6	2	55,8	10,53
Sujeto	628,4	4	157,1	
RESIDUAL	42,4	8	5,3	
TOTAL	782,4	14		

$10.53 \in (F_{\alpha,a-1,(a-1)(b-1)}, \infty) = (F_{0.05,2,8}, \infty) = (4.46, \infty)$. Por tanto, rechazo H_0, sí que hay diferencias entre los tiempos medios (para realizar la tarea) de las tres interfaces.

Es interesante también observar los resultados que se hubiesen obtenido si no se hubiesen considerado los bloques, es decir, vamos a comparar el diseño en bloques con el diseño aleatorizado para mostrar los beneficios de analizar por bloques.

Ejemplo 3.7: Supongamos ahora que no se han considerado los 5 sujetos como bloques, sino que se supone que cada vez se prueba con individuos distintos, es decir, que tuviéramos un diseño completamente aleatorizado. La tabla ANOVA sería:

Fuente	Suma de cuadrados	GL	Cuadrados medios	F
Interfaz	111,6	2	55,8	1,00

```
Error              670,8          12          55,9
------------------------------------------------------------
Total              782,4          14
```

Esta vez $1 \notin (F_{\alpha,a-1,N-a}, \infty) = (F_{0.05,2,12}, \infty) = (3.89, \infty)$. Con lo cual, no rechazaría H_0, no tendría pruebas para afirmar que hubiera diferencia entre los tiempos medios de las 3 interfaces.

Con este ejemplo, vemos pues, que es fundamental el uso de los bloques para poder, en este caso, descubrir la diferencia entre los tiempos medios de las 3 interfaces. El no usar los bloques, nos conduciría a otra conclusión.

De igual forma que antes, es importante examinar los residuos. Ahora los valores valores ajustados o estimados serían: $\hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j = \bar{y}\ \ + (\bar{y}_{i.} - \bar{y})$ $+ (\bar{y}_{.j} - \bar{y})\ \ = \bar{y}_i + \bar{y}_j - \bar{y}$. Una gráfica de los residuos frente a los valores ajustados con forma de curva, podría sugerir una interacción entre los tratamientos y los bloques.

También, siguiendo el esquema del punto anterior, si el análisis de varianzas hubiera indicado la existencia de diferencias entre las medias de los tratamientos, podremos utilizar el método LSD, ahora calculada como LSD = $t_{\alpha/2,(a-1)(b-1)}\sqrt{2CM_E/b}$.

R: multicomp.lm(objeto aov, method="lsd",error.type="cwe")

Ejemplo 3.7: Calcula la LSD al 95 % y realiza las comparaciones explicando los resultados. Identifica grupos homogéneos con columnas de X's. ¿Cuál/es interfaz/s serían mejores?

```
>a=aov(Tiempo~factor(Interfaz)+factor(Sujeto))
> multicomp.lm(a, method="lsd",error.type="cwe")
 $table
    estimate   stderr     lower      upper
1-2     -4.2 1.456022 -7.5575927 -0.8424073
1-3      2.4 1.456022 -0.9575927  5.7575927
2-3      6.6 1.456022  3.2424073  9.9575927
```

Gracias a la salida del R, podemos ver que hay diferencia entre μ_1 y μ_2, puesto que 0 $\notin$ (-7.5575927, -0.8424073). Además, también hay diferencia entre μ_2 y μ_3, puesto que 0 $\notin$ (3.2424073, 9.9575927). Es una manera equivalente a calcular la LSD $= t_{\alpha/2,(a-1)(b-1)}\sqrt{2CM_E/b} = t_{0.025,8}\sqrt{2 \cdot 5.3/5} = 2.306\sqrt{2 \cdot 5.3/5}$ $= 3.3576$ y ver si el valor absoluto de la diferencia de la pareja de medias es superado o no. Por ejemplo, 1-2: $|\bar{y}_{1.} - \bar{y}_{2.}| = |45.6 - 49.8| = 4.2 > 3.3576 \rightarrow$ $\mu_1 \neq \mu_2$. Fíjate que -4.2 $\pm$ 3.3576 = (-7.5575927, -0.8424073).

Estos son los valores de cada media.

```
>model.tables(a,"means")
```

```
Tables of means
Grand mean

46.2

 factor(Interfaz)
   1    2    3
45.6 49.8 43.2

 factor(Sujeto)

    1     2     3     4     5
55.33 49.33 41.00 37.00 48.33


O sea,

Interfaz      Media          Grupos homogéneos
-------------------------------------------------
3             43.2             X

1             45.6             X

2             49.8                X
-------------------------------------------------
```

Los mejores interfaces serían el 3 y el 1, que forman un grupo homogéneo, con menores tiempos medios, que con la interfaz 2.

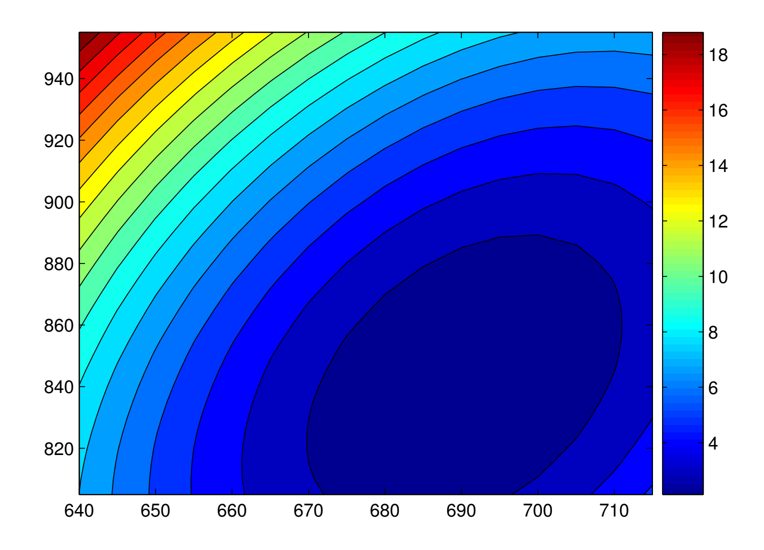
5.4. Diseño factorial con dos factores

El último modelo de ANOVA que se estudiará es el de dos factores, suponemos ahora que la observación de la variable Y está influida por dos factores de interés.

La idea básica de los diseños factoriales es cruzar los niveles de los factores a todas las combinaciones posibles, ya que de esta manera podemos detectar la interacción de factores. Existe interacción entre dos factores, si el efecto de algún nivel de un factor cambia al variar de nivel en el otro factor. En estos experimentos nos interesará estudiar la influencia de los dos factores en la respuesta y la interacción entre los factores. Resulta conveniente resaltar las ventajas de estos diseños frente a experimentos en los que se varía un factor, dejando constantes los demás.

Ejemplo 5.4: Imaginemos que tenemos un proceso para el que queremos minimizar un cierta respuesta R, que viene influenciada de dos variables A y

B. Una posible opción para encontrar unas condiciones óptimas de trabajo en las cuales R sea mínima (y que veremos que no será la adecuada), sería coger el valor de A con el que se trabaja habitualmente (660) y mirar para dicho valor, cuál es el valor de B con el que se obtiene el óptimo, con el que se obtiene menor R, que sería en este caso 815 (con respuesta 4.04). Después, fijado B a 815, variamos A, que nos daría un valor óptimo de 685 (con respuesta 2.28). ¿Es éste el valor óptimo? Pues si miramos la figura siguiente veremos que no (sería 690.6 para A y 841 para B, que nos daría una respuesta de 2.03), que encontraremos si experimentamos con todas las combinaciones de niveles de las dos variables.



Ejemplo 5.5: En la limpieza del hogar, dos productos conocidos son el salfuman y la lejía. El uso más conocido del primero es el de desincrustante para eliminar residuos de caliza. Mientras que el segundo, entre otras cosas, puede usarse para desinfectar los lavabos. Si sus efectos se sumaran, tendríamos el producto de limpieza *Estrella*, pero resulta que si los mezclamos, se produce el tóxico gas cloro (¡cuidado!), vemos que hay interacción.

Ejemplo 5.6: Una ingeniera desea determinar si existe o no diferencia significativa entre los efectos de dos algoritmos (A y B) diseñados para contar olivos en imágenes aéreas, que es de interés para las políticas de subvenciones de la Unión Europea en agricultura. Estos algoritmos, pueden aplicarse a dos tipos de imágenes (con distinta banda espectral y resolución), F y G. La tabla siguiente da las clasificaciones del sistema para cada combinación algoritmo-tipo de imagen.

	Tipo de imagen	
Algoritmo	1	2
A	82	73
	78	69
	84	67
B	63	62
	65	66
	59	67

Para este modelo, consideraremos que tomamos una muestra de tamaño n en cada combinación de factores, uno con a niveles (factor A) y el otro con b (factor B), y las abn observaciones se realizan en orden aleatorio.

Las hipótesis básicas del modelo son ahora:

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk} \quad i = 1, \ldots, a \quad j = 1, \ldots b \quad k = 1, \ldots, n,$$

donde las variables ϵ_{ijk} son $N(0, \sigma^2)$ independientes.

El modelo descompone la respuesta en:

- Una media global μ.
- El efecto incremental en la media debida al nivel i-ésimo del factor A, τ_i ($\sum_i \tau_i = 0$).
- El efecto incremental en la media debido al nivel j-ésimo del factor B, β_j ($\sum_j \beta_j = 0$).
- $(\tau\beta)_{ij}$ representa el efecto de la interacción entre ambos factores, así que $\sum_i (\tau\beta)_{ij} = 0$ y $\sum_j (\tau\beta)_{ij} = 0$.
- El error experimental, ϵ_{ijk}, que recoge el efecto de todas la restantes causas posibles de variabilidad del experimento.

Tres son las hipótesis de interés: $H_0 : \tau_i = 0 \ \forall i$, $H_0 : \beta_j = 0 \ \forall j$ y $H_0 : (\tau\beta)_{ij} = 0 \ \forall i, j$.

Análogamente, el análisis de varianza prueba estas hipótesis descomponiendo la variabilidad total de los datos en sus partes componentes, y comparando después los diferentes elementos de esta descomposición. En la tabla 5.3 podemos ver la tabla de ANOVA para este modelo.

Fuente de variación	Suma de cuadrados	Grados de libertad	Media de cuadrados	F
Factor A	$bn \sum_i (\bar{y}_{i..} - \bar{y}_{...})^2$	$a-1$	$\frac{SC_A}{a-1}$	$\frac{CM_A}{CM_E}$
Factor B	$an \sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2$	$b-1$	$\frac{SC_B}{b-1}$	$\frac{CM_B}{CM_E}$
Interacción	$n \sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$	$(a-1)(b-1)$	$\frac{SC_{AB}}{(a-1)(b-1)}$	$\frac{CM_{AB}}{CM_E}$
Error	$\sum_i \sum_l \sum_k (y_{ijk} - \bar{y}_{ij.})^2$	$ab(n-1)$	$\frac{SC_E}{ab(n-1)}$	
Total	$\sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2$	$abn-1$		

Tabla 5.3: Tabla ANOVA de dos factores con interacción.

R: aov(respuesta ∼ factor1 * factor2)
anova(a)

Ejemplo 5.6: Tabla ANOVA.

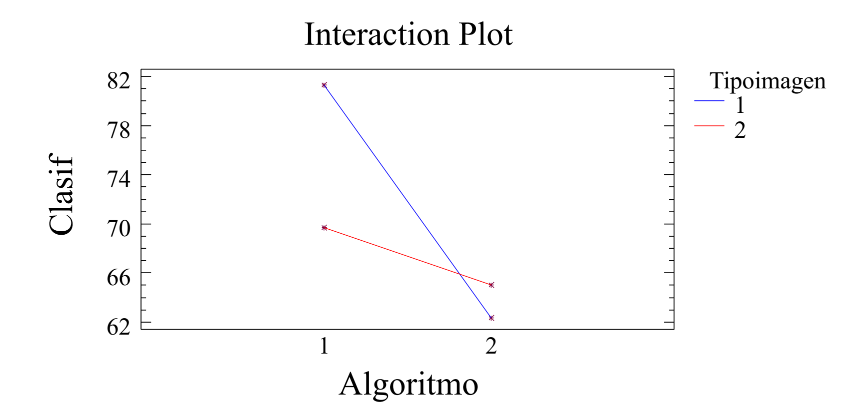
```
Analysis of Variance Table

Response: Clasificacion
                           Df Sum Sq Mean Sq F value    Pr(>F)
factor(Alg)                1 420.08 420.08  48.0095 0.0001210 ***
factor(Tipoi)              1  60.75  60.75   6.9429 0.0299481 *
factor(Alg):factor(Tipoi)1 154.08 154.08  17.6095 0.0030114 **
Residuals                  8  70.00   8.75
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Debemos mirar primero el efecto de la interacción, y posteriormente los efectos principales. Si la interacción no es significativa, podemos analizar los efectos de los factores principales. En cambio, si es significativa, la interpretación de los efectos principales ya no es tan clara, puesto que la interacción puede enmascarar los efectos principales.

En nuestro ejemplo, la interacción y los factores principales son significativos (p-valores menores que $\alpha = 0.05$), con el siguiente gráfico podemos ver claramente la interacción (hay cruce de los segmentos que unen las medias de cada grupo, es decir, tienen un comportamiento diferente según el nivel del otro factor).

Como antes, se deben examinar los residuos: $e_{ijk} = y_{ijk}$ - $\bar{y}_{ij.}$ y también, podemos realizar pruebas para medias individuales, usando el método de la LSD. Si la interacción es significativa, podría aplicarse a las medias de un factor, fijando el otro a un nivel particular.



En el caso en que sólo dispusiéramos de una observación por celda, una sola réplica, hay tantos parámetros en el modelo como observaciones y los grados de libertad del error son cero. Una posible consideración es suponer que el efecto interacción se puede omitir (lo cual no deja de ser peligroso). En este caso, el análisis sería equivalente al usado en el diseño de bloques aleatorizados, aunque debe hacerse notar que las situaciones experimentales que conducen a estos modelos son muy distintas.

PARTE III
APÉNDICE

Capítulo 6

Software

Es indiscutible la importancia del uso de las clases de ordenador para la enseñanza de la Estadística en la actualidad. Aquí únicamente pretendemos recopilar de forma general, herramientas que usaremos en las clases de laboratorio.

En la asignatura anterior a ésta, ya se usó y presentó el R (véase [34] para un repaso). El R es una buena elección porque permite al usuario programar fácilmente funciones adicionales si, llegado el caso, el paquete no contuviera el procedimiento estadístico necesario. El R es apropiado porque facilita la comprensión de los conceptos vistos en teoría, al centrarnos únicamente en los comandos que realizan lo visto en teoría. Muchos programas comerciales, muestran por pantalla una gran cantidad de resultados y menús, mucho mayor que la vista en clase, lo cual hace que para llegar a discernir lo fundamental que se ha visto en clase, se tenga que navegar bastante por distintos menús.

También hemos de tener en cuenta que al tener código abierto, podemos ver lo que se hace en cada instrucción, lo cual es imposible de hacer con un programa comercial, que no permite el acceso al código. Esto convierte al R en un programa más flexible.

Además, otra característica fundamental del R es su carácter gratuito. Recordemos que en los estatutos de la Universitat Jaume I se señala que: «La Universitat Jaume I fomentarà l'ús de formats informàtics oberts en la comunicació interna i externa, promoure el desenvolupament i l'ús del programari lliure i afavorir la lliure difusió del coneixement creat per la comunitat universitària». Pensemos que podemos usar el R en **cualquier plataforma de forma gratuita** y, por tanto, también podéis emplearlo posteriormente en cualquier empresa, sin tener que gastarse miles de euros en un sola licencia.

Por otro lado, el R está en continua expansión, y cuenta con muchísimas librerías que recogen los últimos avances en Estadística, muchos de los cuales no están disponibles en los programas de pago.

Por último, a través de la función *Sweave*() de R [41, 52] de la librería *Hmisc*, podemos fácilmente mezclar texto LaTeXy código R para la generación automática de documentos, es decir, texto y análisis de datos de alta calidad, y ¡¡¡GRATIS!!!, ¿qué más se puede pedir?

6.1. Repaso previo. Simulación y fiabilidad

Objetivos: Aprender a generar números aleatorios de distintas distribuciones. Manejar los conceptos básicos de fiabilidad de componentes. Simular sistemas (sencillos) para verificar su fiabilidad. Repasar conceptos vistos en la asignatura del curso anterior IG12 Estadística.

6.1.1. *Software* de las prácticas

Usaremos R, versión de libre disposición del lenguaje S-PLUS. Es un intérprete de comandos con una gran cantidad de funciones, orientado fundamentalmente al análisis estadístico. Se puede obtener en http://cran.R-project.org/, donde también es posible obtener distinta documentación en inglés o en castellano: http://cran.r-project.org/other-docs.html#nenglish

Es el mismo programa que se utilizó en las prácticas de la asignatura IG12 el curso pasado. Se empleará las funciones básicas y la librería que se encuentra en *Contributed packages: qcc*, la librería de control de calidad. Está disponible tanto en Linux como en Windows.

También es posible obtener el código fuente, en *source code.*

Recuerda que puedes usar help() para ayuda on-line, o help.start() si deseas la ayuda a través de un navegador.

Para salirte de R, teclea 'q()' .

6.1.2. Introducción

El objetivo de la práctica es dar a conocer (muy someramente) dos áreas de gran interés: la simulación y la fiabilidad. Comenzaremos definiendo ambas materias:

La simulación es una potente herramienta para modelizar y analizar sistemas complejos. La mayoría de sistemas reales son difíciles de estudiar a través de modelos analíticos. En cambio, un modelo se simulación puede construirse casi siempre y su ejecución (en el ordenador) genera historias del sistema de las que podemos extraer información estadística. (Nota: también existe la simulación física, ejemplos de la cual aparecen en muchos de los capítulos de la famosa serie de televisión $csi\ LasVegas$).

Los campos de aplicación de la simulación son variadísimos: sistemas de producción/inventario, redes de distribución, sistemas informáticos (sistemas

cliente/servidor, redes de telecomunicaciones, etc.), sistemas de transporte (aeropuertos, puertos, ferrocarriles, autopistas), etc. Algunos ejemplos conocidos de gran renombre son: el proyecto del Eurotúnel o las operaciones del canal de Suez.

Para poder realizar simulaciones, necesitamos saber generar números aleatorios, que nos permitan generar entradas, con las que «alimentar»el modelo durante la simulación. Nosotros nos limitaremos a dar una ligera introducción a la generación de variables aleatorias, ya que la puesta en práctica de un modelo de simulación requiere de números aleatorios (pseudoaletorios), pero no iremos más allá, para ello está la materia optativa correspondiente.

En segundo lugar, se define la *fiabilidad* de un componente (o de un sistema) como la probabilidad de que el componente (o el sistema) funcione en un intervalo de tiempo en condiciones especificadas.

6.1.3. Generación de números aleatorios

Es habitual que los programas para la realización de cálculos estadísticos incorporen un apartado dedicado a la generación de variables aleatorias. Nosotros veremos cómo hacerlo. En este punto, deberíamos empezar considerando cómo generar valores aleatorios de una Uniforme(0,1).

La mayoría de los lenguajes de programación dispone de alguna función para su generación. Debido a la limitación del tiempo, sólo proporcionaremos la siguiente información bibliográfica que puede encontrarse en la red sobre cómo generarlos, y nos restringiremos a utilizar los valores que nos suministren dichas funciones: http://www.library.cornell.edu/nr/bookcpdf/c7-1.pdf (capítulo 7 del libro on-line Numerical Recipes in C) [59] o sobre el generador de números aleatorios de R (http://www.stats.ox.ac.uk/pub/MASS4/VR4stat.pdf).

Sólo veremos cómo generar valores aleatorios de una variable exponencial y Normal. Aunque para generar números aleatorios de una determinada distribución podemos utilizar los comandos disponibles (*r'nombre de la distribución',* por ejemplo: *rbinom, rexp, rnorm, rpois, runif,* etc.) vamos a generarlos a través de la Uniforme(0,1):

runif(n, min=0, max=1)

n	Número de observaciones
min,max	Límites inferior y superior de la distribución

Para cada distribución, el primer argumento indicará el número de observaciones a generar, y los siguientes serán distintos parámetros de las distribuciones, cuyo significado dependerá de la propia distribución.

Sin embargo, antes de comenzar con la generación de la exponencial, haremos un ejercicio previo (¡el calentamiento!) para recordarla.

Actividad 1. La exponencial y su papel en fiabilidad

Para hacer esta actividad, recuerda que para determinar probabilidades de una distribución en el R usamos: (*p'nombre de la distribución',* por ejemplo: *pbinom, pexp, pnorm, ppois, punif,* etc.).

pexp(q, rate = 1, lower.tail = TRUE, log.p = FALSE)

q vector de cuantiles
Media = 1/ rate
lower.tail; si TRUE (por defecto), probabilidad $P[X <= x]$, sino, $P[X > x]$
log; si TRUE, probabilidades p vienen dadas como log(p)

Para cada distribución, el primer argumento indicará el/los valor/es para el que queremos calcular el valor/es de la función de distribución F(x) = P(X $\leq$ x) (con *lower.tail=FALSE* obtendremos su contrario, el área de la cola superior, P(X>x)), y los siguientes serán distintos parámetros de las distribuciones, cuyo significado dependerá de la propia distribución.

El tiempo de duración de un ensamble mecánico en una prueba de vibración tiene una distribución Exponencial con media 400 horas. Calcula y escribe los comandos que utilices:

a) ¿Qué parámetro tendrás que emplear en la función *pexp* como *rate*?

b) ¿Cuál es la probabilidad de que el ensamble falle durante la prueba en menos de 100 horas?

c) ¿Cuál es la probabilidad de que el ensamble trabaje durante más de 500 horas antes de que falle?

d) Si el ensamble se ha probado durante 400 horas sin fallo alguno, ¿cuál es la probabilidad de que falle en las siguientes 100 horas?

En este último apartado acabamos de comprobar la propiedad de falta de memoria de la exponencial.

Actividad 2. Generación de una muestra aleatoria de una distri - bución exponencial. Método de la transformada inversa

Sea F una función de distribución (estrictamente creciente) de una variable aleatoria continua X y U una variable aleatoria uniforme en (0,1). Entonces, $X = F^{-1}(U)$, es una variable aleatoria con distribución F.

Para el caso de la exponencial de parámetro a, tendremos, por tanto: $x =$

-(1/a) log(1-u), o equivalentemente, x = *-(1/a) log(u)* siendo u un valor aleatorio de una variable aleatoria Uniforme(0,1).

2.1. Genera una muestra de tamaño 100, de una exponencial de parámetro 2 mediante este método, es decir, primero genera 100 valores de una Uniforme(0,1) y luego transforma estos valores. Guarda los valores obtenidos, pues se usarán en otra práctica. Incluye los datos generados en la memoria. ¡Fíjate que tus datos serán diferentes a los de tus compañeros!

Puedes usar *write.table(x, file = “ ”)* recuperables con *read.table*, o bien, *save(x,file=“ ”)*, recuperables con *load(file)*.

2.2. Describe los valores obtenidos, incluye en la memoria: el histograma, la media y la varianza. ¿Cuáles eran los valores de la media y varianza de la población de la que hemos generado los valores? Recuerda que para describir una muestra podemos usar: summary(x) que incluye la media (mean(x)), var(x) para la varianza, hist(x) para el histograma, boxplot(x) para el diagrama de cajas, etc. Nota para los gráficos en Linux: x11(), jpeg(“fichero”), hist(x), graphics.off().

Actividad 3. Generación de una muestra aleatoria de una distribución Normal. Recordatorio del teorema central del límite

Para generar valores aleatorios de una Normal(0,1) vamos a utilizar el teorema central del límite, que visteis el curso pasado.

Teorema central del límite: Sean X_1, X_2, ..., X_N variables aleatorias independientes e idénticamente distribuidas tales que $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$, ambas finitas. Entonces cuando N es grande, la variable aleatoria $X = X_1 + X_2 + ... + X_N$ sigue aproximadamente una distribución Normal con media $N\mu$ y varianza $N\sigma^2$.

Vamos a considerar 12 muestras aleatorias independientes de Uniforme(0,1), con lo cual, por el teorema central del límite tendremos $\sum_{i=1}^{12} U_i \approx N(6,1)$, y restándole 6 conseguiríamos una variable $Z \sim N(0,1)$. Para generar $X \sim N(\mu,\sigma^2)$ a partir de Z, basta con invertir el proceso de tipificación: $X = \mu + \sigma Z$.

3.1. Genera una muestra de tamaño 200 de una Normal con media d, siendo d los 4 últimos dígitos de tu DNI y desviación típica 2. Por ejemplo, si tu DNI es: 12345678, entonces *d=5678*. Para ello sigue los pasos siguientes. Genera 12 muestras de tamaño 200 de una Uniforme(0,1), de la siguiente forma:

1. Genera 12 x 200 = 2400 valores de una uniforme(0,1) en un vector llamado *x*.
2. Crea un vector llamado r para codificar las 12 muestras de tamaño 200, es decir, crea un vector del 1 al 200, cada uno de ellos repetido doce veces. Sugerencia: mira la ayuda de *rep*.

3. Seguidamente, realizaremos las sumas. Para ello usaremos la siguiente instrucción: *sumax =aggregate(x,list(r),FUN=sum)*. Mira y copia la ayuda de esta función para asegurarte de lo que hacemos.

4. Para acabar de generar la Normal que queríamos, hemos de restarle 6 e invertir el proceso de tipificado: *d + 2 *(sumax$x - 6),* o sea, si $d =$ 5678, escribiríamos *5678 + 2 *(sumax$x - 6)*. Incluye estos 200 valores en la memoria y guárdalos para una próxima práctica.

3.2. Vamos a comprobar visualmente que los datos anteriores son Normales, con la media y varianza pedidas, para lo cual incluye en la memoria: el histograma, la media y la varianza de esta variable.

Existen otros métodos de generación de variables aleatorias que no se tratarán. En el libro [68], podéis encontrar un amplio tratamiento.

Para finalizar la práctica, vamos a simular sistemas para verificar su fiabilidad.

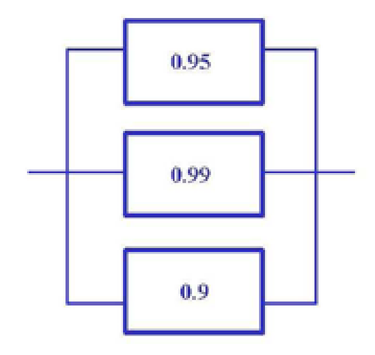
Existen diversas configuraciones: en serie, paralelo, combinaciones de éstos y otros sistemas que no están dispuestos ni en paralelo ni en serie.

Supondremos en lo que sigue que el funcionamiento de cada componente es independiente del de los demás.

Por ejemplo, para un sistema en serie como el siguiente (el sistema funciona si y sólo si todos sus componentes funcionan), la fiabilidad del sistema la calcularíamos como el producto de las fiabilidades de sus componentes.



En una configuración en paralelo como el siguiente, el sistema funciona si, y sólo si, al menos uno de sus componentes funciona, por tanto, deberíamos calcular la probabilidad de la unión. Este cálculo se facilita si calculamos la probabilidad del suceso contrario y usamos las leyes de De Morgan.



También existen sistemas *k de n*. En una configuración *k de n*, el sistema funciona si al menos funcionan k de los n componentes. Nótese que los sistemas en serie y en paralelo son casos particulares de este sistema con $k = n$ y $k = 1$, respectivamente.

Actividad 4. Simulación de sistemas 3 de 5

Vamos a simular el funcionamiento de dos sistemas *3 de 5*, con dos conjuntos de fiabilidades. Pero primero veamos un procedimiento para generar valores de variables aleatorias discretas.

Si tenemos una variable discreta X, que toma valores x_i con probabilidades p_i (recuerda que sumarán 1), un algoritmo para simular X sería: generar valores de una variable $U \sim$ Uniforme(0,1) y hacer $X = x_1$ si $u \leq p_1$, y hacer $X = x_j$ si $\sum_{i=0}^{j-1} p_i < u \leq \sum_{i=0}^{j} p_i$

4.1. Vamos a calcular la fiabilidad de un sistema *3 de 5*, simulando el sistema. La probabilidad de que funcione cada una de las 5 componentes es: 0.9, 0.8, 0.7, 0.6 y 0.5. El siguiente código simula 5 variables, que representan si la componente funciona o no. Así, por ejemplo, para la componente 1, X_1 =1 (funciona) con probabilidad 0.9, y X_1 =0 (no funciona) con probabilidad 0.1. Para cada componente del sistema, generamos 1000000 valores de una Uniforme(0,1), conjuntamente con la indicación de si funciona o no.

```
c1<-runif(1000000)<.9
c2<-runif(1000000)<.8
c3<-runif(1000000)<.7
c4<-runif(1000000)<.6
c5<-runif(1000000)<.5
```

Puedes comprobar, por ejemplo, que si calculamos la media de c1, obtendremos 0.9 aproximadamente. Añade este resultado en la memoria:

```
mean(c1)
```

Para acabar de simular el sistema, sumaremos las variables y veremos si 3 o más componentes funcionan:

```
sumar<-c1+c2+c3+c4+c5
sistema<-sumar>=3
```

Por último, la fiabilidad del sistema, la podemos calcular mediante la media de la variable anterior:

```
mean(sistema)
```

Añade este valor en la memoria.

4.2. Vamos a calcular la fiabilidad de otro sistema *3 de 5.* La probabilidad de que funcione cada una de las 5 componentes es: 0.7. En este caso, Σ X_i sería una Binomial(5,0.7).

a) Vamos a calcular la probabilidad teórica y la obtenida simulando el sistema. Primero simularemos el sistema:

```
c1<-runif(1000000)<.7
c2<-runif(1000000)<.7
c3<-runif(1000000)<.7
c4<-runif(1000000)<.7
c5<-runif(1000000)<.7
sumar<-c1+c2+c3+c4+c5
sistema<-sumar>=3
mean(sistema)
```

Incluye en la memoria, la fiabilidad del sistema obtenida mediante simulación.

b) Ahora calcula la probabilidad teórica (probabilidad de que una variable Binomial(5,0.7) sea mayor o igual que 3) y añádelo en la memoria. Recuerda que *pbinom(q, size, prob, lower.tail = TRUE)* proporciona la función de distribución de una binomial de tamaño *size* y probabilidad de éxito *prob*. Añade este valor a la memoria.

c) Juega (aumenta y disminuye) con el número de simulaciones realizadas anteriormente y comenta lo que sucede.

6.2. Intervalos de confianza y contrastes de hipótesis

Objetivos: Conocer los procedimientos para obtener intervalos de confianza, realizar contrastes paramétricos y no paramétricos. Afianzar los conceptos

tratados en los temas 1 y 2 del módulo teórico, sabiendo interpretar la información que puede extraerse de un intervalo de confianza, y sabiendo plantear y resolver problemas usando los contrastes de hipótesis, comprendiendo los conceptos relativos a los mismos.

6.2.1. Introducción

En esta práctica trabajaremos los intervalos de confianza y contrastes de hipótesis (temas 1 y 2 de teoría). Empezaremos por los intervalos de confianza y contrastes paramétricos para medias, varianzas y proporciones. Después seguiremos con los no paramétricos: test ji-cuadrado (de bondad de ajuste y para tablas de contingencia) y otros tests de bondad de ajuste.

6.2.2. Inferencia paramétrica

Se presentan a continuación, diversos procedimientos que nos permitirán obtener intervalos de confianza y realizar contrastes para medias, varianzas y proporciones.

Medias

t.test(x, y = NULL,alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95, ...)

x	Vector numérico de datos.
y	Vector numérico de datos: OPCIONAL. Sólo lo usaremos si estudiamos dos medias.
alternative	Cadena de caracteres especificando la hipótesis alternativa, que será una de las siguientes opciones: "two.sided"(por defecto), "greater" o "less". Con la letra inicial es suficiente.
mu	Un número indicando el valor verdadero de la media (o la diferencia de medias, si estamos trabajando con dos muestras). Por defecto: 0.
paired	Valor lógico que indica si las muestras son apareadas. Por defecto: FALSE.
var.equal	Variable lógica indicando si consideramos las dos varianzas como iguales, en el caso de dos muestras independientes. Por defecto: FALSE.
conf.level	Nivel de confianza del intervalo (1 - α). Por defecto: 0.95.

Devuelve el contraste e intervalo de confianza para la media o medias, según lo que le hayamos especificado. El p-valor devuelto, nos indicará si rechazar o no la hipótesis nula.

Varianzas

var.test(x, y, ratio = 1,alternative = c("two.sided", "less", "greater"), conf. level = 0.95, ...)

x	Vector numérico de datos.
y	Vector numérico de datos.
ratio	Valor del cociente de varianzas poblacionales de x e y. Por defecto: 1.
alternative	Cadena de caracteres especificando la hipótesis alternativa, que será una de las siguientes opciones: "two.sided"(por defecto), "greater" o "less". Con la letra inicial es suficiente.
conf.level	Nivel de confianza del intervalo (1 - α). Por defecto: 0.95.

Devuelve el contraste e intervalo de confianza para el cociente de varianzas, según los valores de los argumentos que hayamos introducido.

Proporciones

prop.test(x, n, p = NULL,alternative = c("two.sided", "less", "greater"), conf.level = 0.95)

x	Vector con los éxitos.
n	Vector con el número de pruebas, es decir, con los tamaños muestrales.
p	Probabilidad de éxito. Por defecto: NULL, no se considera.
alternative	Cadena de caracteres especificando la hipótesis alternativa, que será una de las siguientes opciones: "two.sided"(por defecto), "greater" o "less". Con la letra inicial es suficiente.
conf.level	Nivel de confianza del intervalo (1 - α). Por defecto: 0.95.

Devuelve el contraste e intervalo de confianza para una o más proporciones, según le hayamos especificado.

Nota: aunque en prácticas usemos esta función que lleva el R en la base (en la librería stats), esta función no devuelve el intervalo visto en teoría, que es el que suele aparecer en los libros de texto, sino el intervalo basado en el estadístico score sin corrección de continuidad, y que según [1] sería preferible. Si quisiéramos obtener el intervalo que calculamos en teoría, que es más sencillo de calcular a mano, tendríamos que usar la función *binconf* de la librería *Hmisc* con la opción "asymptotic".

6.2.3. Inferencia no paramétrica

En esta sección se recogen diferentes procedimientos para contrastar el ajuste a diferentes distribuciones, y el test de la ji-cuadrado para la bondad de ajuste y las pruebas para tablas de contingencia vistas en clase. Empezaremos con este último test.

Test Chi-cuadrado

chisq.test(x, p = rep(1/length(x), length(x)))

x	Vector o matriz de datos.
p	Vector de probabilidades de la misma longitud que x. Por defecto: equiprobables (asume una uniforme discreta).

Si x es un vector, se realiza un test de bondad de ajuste, donde la hipótesis nula sería si las probabilidades poblacionales son iguales a las recogidas en el vector p.

Si x fuera una matriz, se considera como una tabla de contingencia. Chisq.test devolverá el valor observado del estadístico χ^2 y también el p-valor que nos indicará si rechazar o no la hipótesis nula.

Otros contrastes no paramétricos

ks.test(x, y, ...)

x	Vector numérico de datos.
y	Puede ser un vector numérico o una cadena de caracteres con el nombre de la función de distribución.
...	Parámetros de la distribución especificada (con caracteres) por y (no estimados a partir de los datos).

Realiza el test de Kolmogorov-Smirnov.

shapiro.test(x): lleva a cabo el test de Shapiro-Wilks de normalidad, para los datos recogidos en el vector x.

qqnorm(x): permite determinar gráficamente si los datos (recogidos en el vector x) proceden de una normal, según sea el ajuste a la recta, que podemos dibujar con qqline(x).

6.3. Control de calidad

Objetivos: Calcular los gráficos de control de calidad tanto para variables como para atributos tratados en el tema de Control de Calidad del módulo teórico. Interpretar estos gráficos. Estimar la capacidad de un proceso.

6.3.1. Introducción

En esta práctica abordaremos el control estadístico de calidad, que se corresponde con el tema 3 de teoría, en concreto trataremos las gráficas de control y diagramas Pareto. El control de calidad se clasifica en:

1. Control en curso de fabricación (de procesos).
2. Control de recepción y de producto acabado.

El control en curso de fabricación se realiza durante la fabricación del producto, a intervalos fijos de tiempo, y tiene por objeto vigilar el funcionamiento del sistema y recoger información para mejorarlo.

El control de recepción y de producto acabado trata de encontrar una buena manera para decidir si un producto verifica las especificaciones establecidas.

Control de procesos

En todo proceso aparece una cierta variabilidad en la calidad, debida a causas aleatorias o no asignables: variabilidad de la materia prima, la precisión de las máquinas y de los instrumentos de medida, destreza de los operarios, etc. Otras causas no aleatorias o asignables (materias primas defectuosas, desgaste de herramientas, deficiente preparación del operario, etc.) producen ciertos efectos previsibles y definidos. Son pocas y de aparición irregular, pero con grandes efectos. Son eliminables. Diremos que un proceso está en estado de control cuando no le afecta ninguna causa asignable. Un instrumento para determinar si se da o no esta situación son las gráficas de control.

El fundamento teórico de una gráfica de control se basa en la construcción, a partir de los valores de la esperanza μ y la desviación típica σ del modelo teórico de distribución que sigue la característica de calidad considerada, de un intervalo de control (generalmente [μ - 3σ , $\mu + 3\sigma$]). Dentro de este intervalo están casi todos los valores muestrales del proceso, si éste se encuentra bajo control. Las muestras se obtienen a intervalos regulares de tiempo. Un punto que cae fuera de los límites de control, indicaría que el proceso está fuera de control.

El control de calidad se realiza observando en cada elemento:

1. Una característica de calidad medible (longitud, resistencia, contenido de impurezas, etc.) que se compara con un estándar fijado. Es el control por variables (gráficas $\overline{X}$, R, S). La característica se supone distribuida normalmente.

2. Control por atributos:

   a) Un atributo o característica cualitativa que el producto posee o no (correcto o defectuoso, por ejemplo). La característica se supone distribuida según una Binomial. Por tanto, se utilizan las gráficas vistas en el tema 3: p y np.

   b) El número total de defectos. La característica se supone distribuida según una Poisson. Por tanto, se utilizan las gráficas vistas en el tema 3: u y c.

   Para realizar estas gráficas, necesitaremos, en primer lugar cargar la librería "qcc" (Quality Control Charts), mediante: *library(qcc)*.

6.3.2. Gráficas $\overline{X}$ y *R; P; U*

Para realizar cualquiera de estas gráficas, emplearemos la misma instrucción, pero variaremos sus parámetros. De hecho, en este enunciado sólo aparecerán los parámetros más relevantes, aunque si escribís *help(qcc)* tendréis el resto de opciones. Se obtiene también el límite superior (UCL = *upper control limit*) e inferior (LCLLCL = *lower control limit*).

qcc(data, type, sizes, center, std.dev, limits, target, labels, newdata, newsizes, newlabels, nsigmas = 3, confidence.level, plot = TRUE, ...).

data	Un data frame, matriz o vector con los datos observados para la variable a representar. Cada fila del data frame o matriz, y cada valor de un vector, se refiere a una muestra o grupo racional.
type	Cadena de caracteres indicando la gráfica a calcular: Estadístico representado: Descripción de la gráfica "xbar" Media: Medias de una variable continua "S" Desviación típica: Desviaciones típicas de una continua "R" Rango: Rangos de una variable continua "xbar.one" Media: Un dato en cada tiempo de una continua "p" Proporción: Proporción de unidades no conformes "np" Cuenta: Número de unidades defectuosas "c" Cuenta: N^o defectos por unidad "u" Cuenta: N^o medio de defectos por unidad
sizes	Un valor o vector de valores que especifica los tamaños muestrales asociados con cada grupo. Para datos continuos dispuestos en un data frame o una matriz, los tamaños muestrales, se obtienen contando los elementos distintos de NA de cada fila. Para las gráficas "p", "np" y "u"este argumento es necesario.
center	Valor indicando el centro (media) del estadístico.
std.dev	Un valor o vector de valores especificando la desviación(es) típica(s) dentro del grupo del proceso.
limits	Un vector de dos valores indicando los límites de control.
target	Un valor indicando el valor "objetivo" del proceso.
labels	Un vector de caracteres con etiquetas para cada grupo
newdata	Un data frame, matriz o vector, como en data, proporcionando más datos que representar, pero no incluidos en los cálculos.
newsizes	Un vector como el argumento sizes, proporcionando más tamaños muestrales de los nuevos datos a representar, pero no incluidos en los cálculos.
newlabels	Un vector de caracteres con las etiquetas de cada nuevo grupo de los nuevos datos incluidos en *newdata*.
nsigmas	Un valor numérico especificando el número de sigmas que usar para calcular los límites de control. Se ignora si se proporciona el argumento confidence.level.
confidence. level	Un valor numérico entre 0 y 1 indicando el nivel de confianza para el cálculo de los límites de probabilidad.
plot	Valor lógico. Si es TRUE se representa el gráfico de Shewhart.

A lo largo de la práctica, se pueden proporcionar los valores de los parámetros cuando el proceso se encuentra bajo control, o bien tendremos que realizar un estudio previo, como se ha visto en el capítulo 4, descartando los valores fuera de control (una vez estudiadas sus causas) y recalculando los límites.

Aunque los gráficos proporcionen pautas para su interpretación, vosotros mismos podéis interpretar la gráfica. He aquí el recordatorio de teoría sobre la interpretación de los gráficos $\overline{X}$ y R:

1. Puntos fuera de control en $\overline{X}$; R en control: indica un cambio en la media.
2. Puntos fuera de control en $\overline{X}$y en R: indica un cambio en la variabilidad.
3. Rachas: 7 puntos consecutivos por encima o debajo de la media (línea central). Puede indicar (si R está bajo control) cambios en la media (por cambios en la materia prima, el servicio de mantenimiento, etc.).
4. Tendencias: 6 puntos seguidos en sentido creciente o decreciente. Indica la presencia de algún factor que influye gradualmente en el proceso: desgaste de la maquinaria, cambios de temperatura, fatiga (en la gráfica $\overline{X}$); envejecimiento de la maquinaria, mezclas (en R en sentido ascendente); mejora de los operarios o del mantenimiento (en R en sentido descendente).
5. Periodicidades o ciclos: repetición de agrupamientos (sucesión de picos y valles). Indican la presencia de efectos periódicos: temperatura, oscilaciones de corriente (en $\overline{X}$); turnos, acciones de mantenimiento (en R).
6. Inestabilidad: grandes fluctuaciones. Puede indicar un sobreajuste de la máquina, mezcla de materiales, falta de entrenamiento del operario de la máquina.
7. Sobreestabilidad: la variabilidad de las muestras es menor que la esperada (acumulación de puntos en la zona central). Puede que los límites estén mal calculados, que se hayan tomado incorrectamente los datos o que se haya producido un cambio positivo temporal cuya causa debe investigarse.

Para los gráficos P y U, los límites los representa no constantes, en un principio. Para especificarlos constantes, repasa la teoría del capítulo 4.

Recordemos también que, cuando un punto muestral caiga fuera de los límites de control, algunas posibilidades serían (haciendo referencia al gráfico p):

1. El proceso ha variado, aumentando o disminuyendo (según el sentido del valor extremo) el valor de p.
2. El sistema de medición ha cambiado (el inspector o los criterios de medida).
3. Se ha cometido un error al estimar el valor de p en dicha muestra.
4. El proceso no ha variado, pero los límites de control son erróneos.
5. Nada ha cambiado, simplemente un suceso poco frecuente ha ocurrido.

6.3.3. Otros comandos

Otras instrucciones útiles son:

qcc.groups(data, sample)

data	Valores observados.
sample	Indicador de muestra para los datos observados.

Permite agrupar fácilmente los datos, devolviendo una matriz de dimensiones adecuadas, de forma que puedan utilizarse como entrada (*input*) en la función *qcc*.

qcc.options(...): controla y devuelve distintas opciones del paquete qcc.

process.capability(object, spec.limits, target, . . .)

object	Un objeto 'qcc' del tipo "xbar".
spec.limits	Un vector indicando los límites de especificación inferior (LSL) y superior (USL).
target	Un valor especificando el objetivo del proceso. Si falta se usa el valor del objeto 'qcc' si no es NULL, sino el objetivo se toma como el valor medio entre los límites de especificación.

Nos devolverá los índices de la capacidad del proceso (Cp,Cp_k), vistos en teoría.

6.3.4. Diagrama Pareto

El diagrama Pareto es un método gráfico para priorizar problemas o las causas que los producen. Consiste en un diagrama de barras ordenadas según su importancia (cada barra corresponde a uno de los distintos factores). Además, representa 2 escalas: frecuencias absolutas y relativas acumuladas (en %).

También devuelve la tabla de frecuencias (absolutas, acumuladas, relativas y relativas acumuladas). Usaremos *pareto.chart(x), x* contiene los valores.

6.3.5. Gráficas CUSUM

Para obtener gráficas de control de suma acumulada, podemos usar la función cusum(object,. . .), que usa un objeto de la clase *qcc*. La interpretación de estos gráficos se encuentra en el material de teoría.

6.4. Diseño de experimentos

Objetivos: Plantear y resolver problemas reales mediante estas técnicas estadísticas (análisis de la varianza) con el apoyo del R, lo cual es de suma

importancia debido al elevado número de cálculos a realizar. Estudiar la adecuación del modelo. Ser capaz de realizar comparaciones entre las medias.

6.4.1. Introducción

En esta práctica trabajaremos el tema de diseño de experimentos, que se corresponde con el tema 4 de teoría. Empezaremos explicando cómo obtener el análisis de la varianza con un solo factor (diseño completamente aleatorizado), siguiendo con el análisis de la varianza con dos factores sin interacción (diseño por bloques aleatorizados) y terminando con el ANOVA de dos factores con interacción (diseño factorial con dos factores).

6.4.2. Análisis de la varianza con un solo factor

Los siguientes comandos nos servirán también para los otros modelos, modificando la fórmula apropiadamente:

aov(formula,. . .)

La fórmula toma la forma siguiente: respuesta $\sim$ términos donde respuesta es el vector (numérico) respuesta (la variable dependiente sobre la que queremos contrastar la igualdad de medias) y en términos indicaremos los factores de la clase *factor*. En el caso de ANOVA de una vía, términos sólo será el $factor$ con los tratamientos. Si contáramos con más factores (por ejemplo, $first$ y *second*), y escribiéramos en términos: $first$ + *second*, indicaría que consideramos todos los términos del $first$ junto con todos los del *second*, eliminando duplicados. Una especificación de la forma $first$: *second* indicaría el conjunto de términos obtenidos de tomar todas las interacciones de todos los términos en $first$ con todos los términos en *second*. La especificación $first$ ∗ *second* indica el cruce de $first$ y *second*, que sería los mismo que $first$ + *second* + $first$: *second*.

Al objeto devuelto por la función anterior podemos aplicarle distintas funciones:

1. *anova*: obtendremos la clásica tabla ANOVA, estudiada en el capítulo 5.
2. *plot:* devuelve distintos gráficos de diagnósticos como son: valores ajustados vs residuos, plot Q-Q normal de los residuos estandarizados o distancias de Cook. Con el argumento *which* podemos seleccionar los gráficos a representar.
3. *model.tables:* calcula tablas resumen, por ejemplo, podemos obtener una tabla de medias con model.tables(objeto aov,"means").

4. *multicomp.lm:* realiza comparaciones múltiples. Podemos obtener la LSD con: multicomp.lm(objeto aov, method="lsd", error.type="cwe"). Mediante *plot* del objeto devuelto, representaremos los intervalos de cada par. Estas funciones son originarias del S, y vienen incluidas en el fichero cm.R que está en el aulavirtual.

5. *residuals:* obtendremos los residuos.

Para realizar un test de homogeneidad (igualdad) de varianzas, puede usarse el test de Barlett: *bartlett.test (formula)*.

Si queremos obtener un diagrama de cajas de la variable respuesta por cada tratamiento, puede utilizarse *plot(formula)*.

El test de Kruskal-Wallis (*kruskal.test(formula)*) es una alternativa no paramétrica al ANOVA de una vía, que usaremos cuando las hipótesis de normalidad e igualdad de varianzas no se cumplan. El p-valor nos indicará si rechazar o no la hipótesis nula.

6.4.3. Análisis de la varianza con dos factores

En este caso distinguiremos entre con o sin interacción, pues en la formula, tal y como se ha explicado previamente, aparecería con * o + respectivamente.

Podemos utilizar las funciones previamente comentadas para el objeto devuelto. Además, en el caso de interacción, para mostrar las interacciones podemos emplear: *interaction.plot(x.factor, trace.factor, response)*, siendo x.factor, el factor cuyos niveles aparecerán en el eje X, trace.factor es el otro factor y response es la respuesta.

6.5. Regresión

Objetivos: Resolver problemas con el apoyo del R, trabajando los principales conceptos de regresión. Comparar los resultados de varios ajustes y validar las hipótesis.

6.5.1. Modelo lineal

En principio trabajaremos con el modelo lineal con errores, normales, independientes homocedásticos:

$$y_i = \sum_{j=0}^{p} \beta_j x_{ij} + e_i,\ e_i \sim NID(0,\sigma^2)$$

En términos matriciales:

$$\boldsymbol{y} = \mathbf{X}\beta + \boldsymbol{e}$$

donde $\boldsymbol{y}$ es el vector de respuestas, $\mathbf{X}$ es la matriz de diseño y tiene columnas $\boldsymbol{x}_0$, $\boldsymbol{x}_1$,..,$\boldsymbol{x}_p$ de variables independientes o predictoras. Muchas veces $\boldsymbol{x}_0$ es una columna de 1s definiendo un término constante o *intercept.*

Con la función *lm(formula,. . .)* ajustaremos modelos lineales. El objeto *formula* sigue las mismas reglas que ya se comentaron en el apartado anterior, además en *formula* pueden verse más detalles ($help(formula)$).

Al objeto devuelto por esta función podemos aplicarle distintas funciones:

1. *summary:* entre otros resultados devuelve un resumen de los residuos, las estimaciones del modelo y los contrastes sobre los coeficientes de regresión, estadístico F o el coeficiente de determinación (corregido), que nos puede proporcionar una primera idea de la bondad del ajuste. Con *anova.lm* obtendremos la tabla ANOVA.

2. *predict.lm:* predice valores basados en el objeto del modelo lineal.

   predict(object,newdata,interval = c(“none”, “confidence”, “prediction”), . . .), según seleccionemos *interval* obtendremos distintos intervalos: intervalo de confianza para la media y los límites del intervalo de predicción.

3. Otras funciones útiles son: *plot.lm* que devuelve distintos gráficos de diagnóstico o *lm.influence* en el mismo sentido, *residuals* para obtener los residuos, *step* para seleccionar el modelo.

Si deseamos representar cada término frente a la variable repuesta podemos usar *plot(formula),* además en el caso de la regresión simple con *abline(objeto lm)*, tendremos la recta ajustada.

Capítulo 7

Formulario

Datos: $\{x_1, x_2, ..., x_N\}$

Media: $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$

Rango intercuartílico: Diferencia entre el tercer y primer cuartil

Varianza: $s^2 = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N-1} = \frac{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}{N-1}$

Desviación típica: s $= \sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N-1}} = \sqrt{\frac{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}{N-1}}$

DISTRIBUCIONES DISCRETAS:

Binomial(n, p):

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, ..., n, \quad q = 1 - p$$

$\mu = n \cdot p$, y $\sigma^2 = n \cdot p \cdot q$

$$\binom{n}{x} = \frac{n!}{x! \cdot (n-x)!} \qquad \text{siendo} \quad n! = n \cdot (n-1) \cdot (n-2) \cdot ... \cdot 2 \cdot 1$$

Poisson(λ):

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0,\ 1,\ 2,\ 3,\ ... \qquad (x \in \mathbb{N})$$

$\mu = \lambda$ y $\sigma^2 = \lambda$

ESTIMACIÓN

Estimador puntual de p: $\frac{X}{N}$, donde X es el número de éxitos en los N experimentos.

Estimador puntual de μ: $\overline{X}$

Estimador puntual de σ^2: $S^2 = \frac{\sum_{i=1}^{N}(X_i-\overline{X})^2}{N-1}$

Estimador puntual del parámetro λ de una Poisson: $\hat{\lambda} = \overline{X}$.

INTERVALOS DE CONFIANZA: tamaño muestral = N, nivel de significación = α

- A) Intervalo de confianza para μ, con σ^2 conocida: ($\overline{x}$ - $z_{\alpha/2}\frac{\sigma}{\sqrt{N}}$, $\overline{x}$ + $z_{\alpha/2}\frac{\sigma}{\sqrt{N}}$) con P(Z $\geq z_{\alpha/2}$) = $\alpha/2$, $Z \sim$ N(0,1)
- B) Intervalo de confianza para μ, con σ^2 desconocida, para Normales:
  ($\overline{x}$ - $t_{\alpha/2}\frac{s}{\sqrt{N}}$, $\overline{x}$ + $t_{\alpha/2}\frac{s}{\sqrt{N}}$) con P(T $\geq t_{\alpha/2}$) = $\alpha/2$, T es t-Student con $N-1$ grados de libertad
- C) Intervalo de confianza para μ, con σ^2 desconocida y N grande ($N \geq$ 30):
  ($\overline{x}$ - $z_{\alpha/2}\frac{s}{\sqrt{N}}$, $\overline{x}$ + $z_{\alpha/2}\frac{s}{\sqrt{N}}$) con P(Z $\geq z_{\alpha/2}$) = $\alpha/2$, $Z \sim$ N(0,1)
- Selección del tamaño de la muestra (media): $N = (\frac{z_{\alpha/2}\cdot\sigma}{Error})^2$
- D) Intervalo de confianza para la diferencia de medias μ_1 - μ_2, con σ_1^2 y σ_2^2 conocidas, para muestras aleatorias independientes (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):
  ($\overline{x}_1$ - $\overline{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{N_1}+\frac{\sigma_2^2}{N_2}}$) con P($Z \geq z_{\alpha/2}$) = $\alpha/2$, $Z \sim$ N(0,1)
- E) Intervalo de confianza para la diferencia de medias μ_1 - μ_2, con σ_1^2 y σ_2^2 desconocidas, para muestras aleatorias independientes y tamaños muestrales grandes (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):
  ($\overline{x}_1$ - $\overline{x}_2 \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}}$) con P($Z \geq z_{\alpha/2}$) = $\alpha/2$, $Z \sim$ N(0,1)
- F) Intervalo de confianza para la diferencia de medias μ_1 - μ_2 de poblaciones normales independientes, con varianzas desconocidas pero iguales ($\sigma_1^2 = \sigma_2^2$) (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):
  ($\overline{x}_1$ - $\overline{x}_2 \pm t_{\alpha/2}\sqrt{\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{N_1+N_2-2}}\sqrt{\frac{N_1+N_2}{N_1N_2}}$) con P($T \geq t_{\alpha/2}$) = $\alpha/2$, T es t-Student con $N_1 + N_2 - 2$ grados de libertad
- G) Intervalo de confianza para la diferencia de medias μ_1 - μ_2 de poblaciones normales independientes, con varianzas σ_1^2, σ_2^2 desconocidas y desiguales (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

$(\overline{x}_1 - \overline{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}})$ con $P(T \geq t_{\alpha/2}) = \alpha/2$, T es t-Student con $\frac{(\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2})^2}{\frac{(s_1^2/N_1)^2}{N_1-1}+\frac{(s_2^2/N_2)^2}{N_2-1}}$ grados de libertad

- H) Intervalo de confianza para la diferencia de medias para muestras apareadas, con diferencia normal:

  $(\overline{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{N}})$ donde $\overline{d}$ es la media de las diferencias y s_d es la desviación típica de las diferencias. Además, $P(T \geq t_{\alpha/2}) = \alpha/2$, T es t-Student con N - 1 grados de libertad, N es el número de objetos (parejas) de que disponemos.

- I) Intervalo de confianza para σ^2 en una población normal:

  $(\frac{(N-1)s^2}{\chi^2_{\alpha/2}}, \frac{(N-1)s^2}{\chi^2_{1-\alpha/2}})$ con $P(\chi^2 > \chi^2_{\alpha/2}) = \alpha/2$, χ^2 es chi- cuadrado con $N-1$ grados de libertad.

- J) Intervalo de confianza para el cociente σ_1^2/σ_2^2 de varianzas de dos poblaciones normales independientes:

  $(\frac{s_1^2}{s_2^2}\frac{1}{F_{\alpha/2}}, \frac{s_1^2}{s_2^2}\frac{1}{F_{1-\alpha/2}})$ donde P(F $> F_{\alpha/2}) = \alpha/2$ y F es F de Snedecor con $(N_1 - 1, N_2 - 1)$ grados de libertad.

- K) Intervalo de confianza para una proporción p (de una Binomial) cuando N es grande y la proporción no es cercana a cero o uno:

  $(\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{N}})$, donde P($Z > z_{\alpha/2}) = \alpha/2$ $Z \sim$ N(0,1) y $\hat{p}$ = X /N, $\hat{q}$ = 1 - $\hat{p}$, X = número de éxitos.

- Selección del tamaño de la muestra (proporción):

  $N = p(1-p) \cdot (\frac{z_{\alpha/2}}{E})^2 \leq \frac{1}{4} (\frac{z_{\alpha/2}}{E})^2$

- L) Intervalo de confianza para una proporción p, si ésta es muy cercana a cero:

  $(0, \frac{1}{2N}\chi^2_\alpha)$ con $P(\chi^2 > \chi^2_\alpha) = \alpha$, χ^2 es chi- cuadrado con $2(X+1)$ grados de libertad, X = número de éxitos

- M) Intervalo de confianza para la diferencia de dos proporciones, con N_1 y N_2 grandes (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

  $(\hat{p_1} - \hat{p_2} \pm z_{\alpha/2}\sqrt{\frac{\hat{p_1}\hat{q_1}}{N_1} + \frac{\hat{p_2}\hat{q_2}}{N_2}})$, donde P($Z > z_{\alpha/2}) = \alpha/2$ $Z \sim$ N(0,1), $\hat{p_1}$ = X_1 /N_1, $\hat{q_1}$ = 1 - $\hat{p_1}$, X_1 = número de éxitos en las N_1 pruebas y $\hat{p_2}$ = X_2 /N_2, $\hat{q_2}$ = 1 - $\hat{p_2}$, X_2 = número de éxitos en las N_2 pruebas.

CONTRASTE DE HIPÓTESIS: tamaño muestral = N, nivel de significación = α

H_1	Región crítica
$\mu < \mu_0$	$(-\infty, -z_\alpha)$
$\mu \neq \mu_0$	$(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$
$\mu > \mu_0$	(z_α, ∞)

- A) Contraste de hipótesis para μ, con N grande:

  $Z = \frac{\bar{X}-\mu_0}{S/\sqrt{N}} \sim$ N(0,1) $H_0 : \mu = \mu_0$

- B) Contraste de hipótesis para μ, con σ^2 desconocida para una población Normal:

  $T = \frac{\bar{X}-\mu_0}{S/\sqrt{N}} \sim t_{N-1}$ $H_0 : \mu = \mu_0$

H_1	Región crítica
$\mu < \mu_0$	$(-\infty, -t_\alpha)$
$\mu \neq \mu_0$	$(-\infty, -t_{\alpha/2}) \cup (t_{\alpha/2}, \infty)$
$\mu > \mu_0$	(t_α, ∞)

- C) Contraste para la diferencia de medias μ_1 - μ_2, con σ_1^2 y σ_2^2 desconocidas, para muestras aleatorias independientes y tamaños muestrales grandes (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

$$Z \approx \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{s_1^2/N_1 + s_2^2/N_2}} \sim N(0,1) \qquad \begin{cases} H_0 : \mu_1 - \mu_2 = \Delta_0 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$\mu_1 - \mu_2 < \Delta_0$	$(-\infty, -z_\alpha)$
$\mu_1 - \mu_2 \neq \Delta_0$	$(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$
$\mu_1 - \mu_2 > \Delta_0$	(z_α, ∞)

- D) Contraste para la diferencia de medias μ_1 - μ_2 de poblaciones normales independientes, con varianzas poblacionales desconocidas pero iguales ($\sigma_1^2 = \sigma_2^2$) (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

$$T = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{N_1+N_2-2}}} \sqrt{\frac{N_1 \cdot N_2}{N_1 + N_2}} \sim t_{N_1+N_2-2}$$

$$\begin{cases} H_0 : \mu_1 - \mu_2 = \Delta_0 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$\mu_1 - \mu_2 < \Delta_0$	$(-\infty, -t_\alpha)$
$\mu_1 - \mu_2 \neq \Delta_0$	$(-\infty, -t_{\alpha/2}) \cup (t_{\alpha/2}, \infty)$
$\mu_1 - \mu_2 > \Delta_0$	(t_α, ∞)

- E) Contraste para la diferencia de medias μ_1 - μ_2 de poblaciones normales independientes, con varianzas poblacionales σ_1^2, σ_2^2 desconocidas y desiguales (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):

$$T = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{s_1^2/N_1 + s_2^2/N_2}} \sim t_{\text{g.l.}}$$

$$g.l. = \frac{(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2})^2}{\frac{(s_1^2/N_1)^2}{N_1-1} + \frac{(s_2^2/N_2)^2}{N_2-1}} \qquad \begin{cases} H_0 : \mu_1 - \mu_2 = \Delta_0 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$\mu_1 - \mu_2 < \Delta_0$	$(-\infty, -t_\alpha)$
$\mu_1 - \mu_2 \neq \Delta_0$	$(-\infty, -t_{\alpha/2}) \cup (t_{\alpha/2}, \infty)$
$\mu_1 - \mu_2 > \Delta_0$	(t_α, ∞)

- F) Contraste para la diferencia de medias μ_D para muestras apareadas, cuya diferencia es normal: $\bar{D}$ y S_D son la media y desviación típica de las diferencias:

$$T = \frac{\bar{D} - \Delta_0}{S_D/\sqrt{N}} \sim t_{N-1} \qquad \begin{cases} H_0 : \mu_D = \Delta_0 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$\mu_D < \Delta_0$	$(-\infty, -t_\alpha)$
$\mu_D \neq \Delta_0$	$(-\infty, -t_{\alpha/2}) \cup (t_{\alpha/2}, \infty)$
$\mu_D > \Delta_0$	(t_α, ∞)

- G) Contraste para σ^2 en una población normal:

$$\chi_0^2 = \frac{(N-1)S^2}{\sigma_0^2} \sim \chi_{N-1}^2 \qquad \begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$\sigma^2 < \sigma_0^2$	$(0, \chi_{1-\alpha}^2)$
$\sigma^2 \neq \sigma_0^2$	$(0, \chi_{1-\alpha/2}^2) \cup (\chi_{\alpha/2}^2, \infty)$
$\sigma^2 > \sigma_0^2$	(χ_α^2, ∞)

- H) Contraste para el cociente σ_1^2/σ_2^2 de varianzas de dos poblaciones normales independientes:

$$F = \frac{S_1^2}{S_2^2} \sim F_{(N_1-1,N_2-1)} \qquad \begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$\sigma_1^2 < \sigma_2^2$	$(0, F_{1-\alpha}) = (0, \frac{1}{F_\alpha^{(N_2-1,N_1-1)}})$
$\sigma_1^2 \neq \sigma_2^2$	$(0, F_{1-\alpha/2}) \cup (F_{\alpha/2}, \infty)$
$\sigma_1^2 > \sigma_2^2$	(F_α, ∞)

- I) Contraste para una proporción p (de una Binomial) cuando N es grande y la proporción no es cercana a cero ni a uno:
$\hat{p} = X/N$ (X = número de éxitos en las N pruebas), q_0 = 1 - p_0

$$Z \approx \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / N}} \sim N(0,1) \qquad \begin{cases} H_0 : p = p_0 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$p < p_0$	$(-\infty, -z_\alpha)$
$p \neq p_0$	$(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$
$p > p_0$	(z_α, ∞)

- J) Contraste para la diferencia de dos proporciones, con N_1 y N_2 grandes (N_1 = tamaño muestral de la muestra de la población 1, N_2 = tamaño muestral de la muestra de la población 2):
$\hat{p_1} = X_1/N_1$ (X_1 = número de éxitos en las N_1 pruebas), $\hat{p_2} = X_2/N_2$ (X_2 = número de éxitos en las N_2 pruebas), $\hat{p} = (X_1 + X_2)/(N_1 + N_2)$

$$Z \approx \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})(1/N_1 + 1/N_2)}} \sim N(0,1) \qquad \begin{cases} H_0 : p_1 = p_2 \\ H_1 : 3 \text{ casos posibles} \rightarrow \end{cases}$$

H_1	Región crítica
$p_1 < p_2$	$(-\infty, -z_\alpha)$
$p_1 \neq p_2$	$(-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, \infty)$
$p_1 > p_2$	(z_α, ∞)

- K) Prueba de la bondad de ajuste con la χ^2:

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i},$$

Bajo H_0, sigue aproximadamente una distribución χ^2 con $k-r-1$ grados de libertad, siendo r el número de parámetros estimados por máxima verosimilitud. La región crítica (a nivel α) es: (χ^2_α, ∞).

- L) Pruebas con tablas de contingencia:

X \ Y	y_1	...	y_j	...	y_c	Total
x_1	o_{11}	...	o_{1j}	...	o_{1c}	$T_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_i	o_{i1}	...	o_{ij}	...	o_{ic}	$T_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_r	o_{r1}	...	o_{rj}	...	o_{rc}	$T_{r.}$
Total	$T_{.1}$	...	$T_{.j}$	...	$T_{.c}$	T

$T_{i.}$ es el total de observaciones de la fila i-ésima, $T_{.j}$ es el total de observaciones de la columna j-ésima y T es el total de observaciones.

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

siendo $e_{ij} = T_{i.} \cdot T_{.j} \ / \ \mathrm{T}$

Bajo H_0, sigue aproximadamente una distribución χ^2 con $(r-1)\cdot(c-1)$ grados de libertad. La región crítica (a nivel α) es: (χ^2_α, ∞).

CONTROL DE CALIDAD

- Gráfico de control $\bar{X}$:

$$\begin{aligned} LSC &= \bar{\bar{x}} + A_2\bar{r} \\ LC &= \bar{\bar{x}} \\ LIC &= \bar{\bar{x}} - A_2\bar{r} \end{aligned}$$

donde $\bar{\bar{x}} = \frac{1}{m}\sum_{i=1}^{m} \bar{x}_i$ ($\bar{x}_i$ es la media muestral de la muestra i-ésima, calculada con los n valores de cada muestra y m es el número total de muestras), $\bar{r} = \frac{1}{m}\sum_{i=1}^{m} r_i$ (donde r_i es el rango de la muestra i-ésima) y la constante A_2 aparece tabulada.

- Gráfico R:

$$\begin{aligned} LSC &= D_4\bar{r} \\ LC &= \bar{r} \\ LIC &= D_3\bar{r}. \end{aligned}$$

Los valores de D_3 y D_4 para distintos valores de n aparecen tabulados.

- Un estimador de σ es $\hat{\sigma} = \bar{R} \ /d_2$, donde d_2 está tabulada.

- Índices de capacidad del proceso:

$$ICP = \frac{LSE - LIE}{6\sigma},$$

donde LSE y LIE son los límites superior e inferior de especificación.

$$ICP_k = \text{mín}\{\frac{LSE - \mu}{3\sigma}, \frac{\mu - LIE}{3\sigma}\}.$$

- Longitud de racha media (ARL):

ARL $= 1/p$, p es la probabilidad de que cualquier punto exceda los límites de control.

- Gráfica P:

$$\begin{aligned} LSC &= \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \\ LC &= \bar{p} \\ LIC &= \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}, \end{aligned}$$

donde $\bar{p}$ es la estimación de p (fracción defectuosa del proceso), obtenido mediante:

$$\bar{p} = \frac{1}{m}\sum_{i=1}^{m} \hat{p}_i$$

con $\hat{p}_i$ la proporción muestral de unidades defectuosas en la muestra i-ésima.

- Gráfico U:

$$\begin{aligned} LSC &= \bar{u} + 3\sqrt{\frac{\bar{u}}{n}} \\ LC &= \bar{u} \\ LIC &= \bar{u} - 3\sqrt{\frac{\bar{u}}{n}} \end{aligned}$$

donde, si tenemos n (que puede no ser un entero) unidades y un total de defectos C entonces:

$$U = \frac{C}{n},$$

es el promedio de defectos por unidad. Con m muestras preliminares y valores aleatorios $U_1, \ldots, U_m$ entonces el número medio de defectos por unidad es:

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m} U_i.$$

DISEÑO DE EXPERIMENTOS:

- **Diseño completamente aleatorizado: análisis de la varianza con un solo factor**

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2),$$

con τ_i definida como desviaciones de la media global μ, por lo que $\sum_{i=1}^{a} \tau_i = 0$.

Denotaremos por n_i las observaciones en el tratamiento i-ésimo y N el total de observaciones, a es el número de niveles del factor.

Fuente de variación	Suma de cuadrados	Grados de libertad	Media de cuadrados	F
Tratamientos (entre grupos)	$\sum_{i=1}^{a} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	$a-1$	$\frac{SC_{Trat.}}{(a-1)}$	$\frac{CM_{Trat.}}{CM_E}$
Error (dentro grupos)	$\sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$	$N-a$	$\frac{SC_E}{N-a}$	
Total	$\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$	$N-1$		

Tabla 7.1: Tabla ANOVA de un factor

Región crítica (a nivel α): $(F_{\alpha,a-1,N-a}, \infty)$

Método de la mínima diferencia significativa o LSD (*Least Significant Difference*): el par de medias μ_i y μ_j se declarará significativamente diferente si $|\bar{y}_{i.} - \bar{y}_{j.}| >$ LSD, donde LSD al nivel α viene definida como:

$$t_{\alpha/2,N-a}\sqrt{CM_E(1/n_i + 1/n_j)}$$

- **Diseño en bloques aleatorizados**

$$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \quad i = 1, \ldots, a \quad j = 1, \ldots b,$$

donde ϵ_{ij} son variables $N(0, \sigma^2)$ independientes, y $\sum_i \tau_i = 0$ y $\sum_j \beta_j = 0$

Fuente de variación	Suma de cuadrados	Grados de libertad	Media de cuadrados	F
Tratamientos	$SC_{Tratamientos}$	$a-1$	$\frac{SC_{Tratamientos}}{(a-1)}$	$\frac{CM_{Tratamientos}}{CM_E}$
Bloques	$SC_{Bloques}$	$b-1$	$\frac{SC_{Bloques}}{b-1}$	
Error	SC_E	$(a-1)(b-1)$	$\frac{SC_E}{(a-1)(b-1)}$	
Total	SC_T	$ab-1$		

Tabla 7.2: Tabla ANOVA de un diseño en bloques aleatorizados

Región crítica (a nivel α): $(F_{\alpha,a-1,(a-1)(b-1)}, \infty)$

Método LSD: LSD $= t_{\alpha/2,(a-1)(b-1)}\sqrt{2CM_E/b}$.

FUNCIONES del R:

- runif(n, min=0, max=1)
- pexp(q, rate = 1, lower.tail = TRUE, log.p = FALSE)
- t.test(x, y = NULL,alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95, ...)
- var.test(x, y, ratio = 1,alternative = c("two.sided", "less", "greater"), conf.level = 0.95, ...)
- prop.test(x, n, p = NULL,alternative = c("two.sided", "less", "greater"),conf.level = 0.95)
- chisq.test(x, p = rep(1/length(x), length(x)))
- ks.test(x, y, ...)
- shapiro.test(x)
- qqnorm(x) qqline(x)
- qcc(data, type, sizes, center, std.dev, limits, target, labels, newdata, newsizes, newlabels, nsigmas = 3, confidence.level, plot = TRUE, ...)
- qcc.groups(data, sample) qcc.options(...)
- process.capability(object, spec.limits, target, ...)
- pareto.chart(x)
- cusum(object, ...)

- aov(respuesta ∼ términos, ...)
- anova plot model.tables multicomp.lm(objeto aov, method="lsd", error.type="cwe") residuals
- bartlett.test (formula)
- kruskal.test(formula)
- interaction.plot(x.factor, trace.factor, response)
- lm(formula, ...)

PARTE IV
BIBLIOGRAFÍA

Capítulo 8

Material bibliográfico

Lo vamos a dividir en tres apartados. En primer lugar, repasaremos aquellos libros que por contenidos y extensión se adecúan casi completamente a la asignatura (podrían ser utilizados como libros de texto en esta asignatura). En segundo lugar, presentaremos los libros que se pueden recomendar a los estudiantes que estén interesados en profundizar en alguno de los temas de esta asignatura y también otras referencias que por ser más amplias o bien por no estar concebidas para ingenieros, pese a que serían excelentes refuerzos para la materia, las calificaríamos como bibliografía complementaria. Y en último lugar, haremos un repaso por diverso material y recursos on-line.

8.1. Bibliografía básica

Aunque muchos otros podrían ser recomendados, se han escogido los siguientes por estar dirigidos casi la totalidad de ellos a estudiantes de Ingenierías. Aparecen ordenados alfabéticamente:

Ardanuy y Martín [5]. Muy ajustado a los contenidos de la asignatura, aunque contiene algunos pocos temas más, como las series temporales y los números índices.

Canavos [10]. Es un libro general, claro y enfocado principalmente a la práctica.

Cao *et al.* [11] y **Vilar** [79]. La obra de Cao *et al.* sería excelente para la mayor parte del programa, ya que incide en la interpretación y la aplicación de los métodos estadísticos más que en la formulación matemática, puesto que está orientada hacia las titulaciones de carácter técnico y experimental, pero a la vez es muy completo. Tiene gran cantidad de ejercicios resueltos y propuestos, muchos de ellos aplicados a la informática. También, incluye colecciones de cuestiones de respuesta múltiple para la autoevaluación de los estudiantes y ejercicios globales que recogen todos los temas estudiados en el libro. Sin embargo, este libro no contiene ningún tema dedicado a modelos lineales, ni a control de calidad. Un libro en la línea del de Cao *et al.*, pero dedicado a los modelos lineales en exclusiva y cuyo autor también pertenece a la Universidad de A Coruña

sería el de Vilar [79], junto con el material que se encuentra en su página *web*: http://www.udc.es/dep/mate/estadistica2/estadistica_2.htm.

Coronado *et al.* [17]. Este libro incluye todos los temas tratados, entrelazando teoría con práctica (con Statgraphics pero en versión MS-DOS), lo cual lo convierte en una de las referencias destacadas. No obstante, no incluye problemas.

Chatfield [13]. Obra general para ingenieros, clara y bastante concisa, con muchos ejemplos.

Devore [22]. Es un libro de estadística para ingenieros. Muchos de los ejemplos y ejercicios involucran aplicaciones a las ciencias biológicas y de la vida.

Domingo [23]. Es un libro muy ajustado a la asignatura, pues su origen son unos apuntes de una asignatura semestral para una ingeniería. No aparece la parte de control de calidad, pese a ello resulta ser una obra que se acopla a la perfección a la asignatura. Está escrito en catalán.

Dougherty [24]. Excelente libro con muchas aplicaciones a la Informática y a la vez con un respetable nivel matemático. Cuenta también con numerosos problemas, cuyas soluciones aparecen al final del texto y hasta 350 problemas resueltos. Recoge todos los contenidos de la asignatura y algunos otros que no se incluye en el programa por falta de tiempo, pero que también son de interés para los ingenieros informáticos. Contiene una pequeña introducción a los siguientes tópicos: cadenas de Markov y entropía. Además, los contenidos sobre diseño de experimentos, y estadística no paramétrica están más ampliados. También cuenta con un apéndice dedicado a revisar el uso de diversos paquetes comerciales de software estadístico como son: SAS, MINITAB y SPSS. Por otro lado, el autor del libro a nivel investigador goza de un gran prestigio en el campo del análisis de imágenes, y en concreto en el área de la morfología matemática.

Fernández *et al.* [29]. Este libro contiene los aspectos que se cubrirán en las prácticas, pero con el *Statgraphics*.

García et al. [31]. Es el libro usado para el autoaprendizaje de la asignatura de Estadística I de Ingeniería Técnica en Informática de Sistemas en la UNED. Tiene una excelente colección de problemas resueltos relacionados con la informática pero no cubre el bloque dedicado a modelos lineales. Otros libros en la misma línea por ser publicaciones de la UNED serían: Hernández *et al.* [36].

Johnson [39]. Libro de estadística para ingeniería, con excelentes ejemplos orientados a la ingeniería y ejercicios propuestos con sus soluciones.

Mendenhall y Sincich [45]. Es un libro orientado a la ingeniería, muy extenso y con muchos ejercicios, muchos de ellos con datos reales y con aplicaciones a la informática.

Montgomery y Runger [49]. Libro de estadística para ingenieros, con una presentación muy sencilla y clara de los conceptos. Tiene un gran número de ejemplos del mundo de la informática.

Nicolás [53]. Libro introductorio que utiliza como *software* estadístico el R.

Peña [56, 55]. Excelente libro que se ha convertido en clásico en la estadística aplicada. Consta de dos volúmenes, en el primero se tratan los tres primeros temas de la asignatura y en el segundo volumen, el bloque de modelos lineales, además de las series temporales que no se incluye en el programa. Los contenidos son un poco más amplios y con un cierto nivel.

Pérez [57]. Es un excelente libro dedicado a mostrar las utilidades del programa comercial Statgraphics, muy completo, con muchos casos prácticos resueltos junto con su interpretación. Trata muchos más tópicos del considerado en el programa.

http://www.r-project.org [60]. En la página web del R, además del programa podemos encontrar distinta documentación muy útil, tanto en inglés como en castellano, y por supuesto, los manuales de R incluidos en todas las instalaciones (ver el directorio *./doc/manual*).

Ras [62]. Está escrito en catalán como un texto básico semestral en una ingeniería. No da ninguna colección de problemas y tampoco trata todos los puntos del programa.

Ras et al. [63]. El libro lo forman siete prácticas resueltas empleando el Statgraphics. Este libro destaca porque las prácticas no se limitan a tratar un tema concreto, sino que son prácticas muy completas que engloban diversos temas del programa.

Romero y Zúnica [67]. Libro enfocado a la docencia de la estadística en las ingenierías, con ejercicios resueltos y propuestos, dando gran importancia al diseño de experimentos y modelos de regresión. En un fascículo independiente se hallan los temas de introducción a los procesos estocásticos y la teoría de colas.

Scheaffer y McClave [69]. Está dedicado a la estadística para la ingeniería, tiene muchos ejercicios.

Ugarte y Militino [75]. Este libro cubre casi todos los temas, a excepción del control de calidad, a un nivel adecuado para ser usado como texto docente. Tiene una orientación práctica, con ejercicios resueltos con S-PLUS y sin este paquete estadístico. Todo ello, lo convierte en un buen texto docente.

Walpole *et al.* [80]. Excelente texto, amplio y muy claro, orientado a ingenieros y útil como referencia.

Durante los últimos años han ido apareciendo diversos libros dedicados a estadística para ingenieros de autores españoles, que se ajustan en mayor o menor medida al programa de la asignatura, ya que son libros orientados a la Estadística en las Ingenierías (tal y como se estructura en la universidad española) en general, pero no a las Ingenierías Informáticas en particular. En la mayoría de estos libros, se presenta los contenidos teóricos fundamentales y numerosos problemas, resueltos y propuestos. A continuación citamos alguno de ellos:

- A. Gámez y L. M. Marín [30].
- M. A. Castro y Y. Villacampa [12].
- F. J. Alonso *et al*. [4].
- J. M. Egusquiza [25].

8.2. Bibliografía complementaria

A continuación se muestra una serie de libros (y otro material) que pueden ser recomendados a los estudiantes para ampliar determinados temas. También incluimos referencias que por ser más amplias o bien por no estar concebidas para ingenieros, pese a que serían excelentes refuerzos para la materia, no las consideraríamos estrictamente como libros de texto básicos para esta asignatura. Sin perjuicio de que en general no las recomendaramos como texto básico, determinados temas de estas referencias más amplias o no dirigidas a las ingenierías, sí que podrían servir como referencias básicas, pues a veces las fronteras entre lo básico y lo complementario no son completamente claras y definidas. En primer lugar, presentaremos en orden alfabético estas referencias generales, y por último, detallaremos otras referencias especializadas en temas más específicos.

Agresti *et al*. [2]. Este libro contiene ejemplos y ejercicios muy interesantes, con explicaciones claras.

Allen [3]. Es un libro muy completo, con un alto nivel y con una excelente orientación a la computación debida a la experiencia del autor como profesor en *Los Angeles IBM Information System Management Institute*. Contiene un excelente capítulo sobre teoría de colas. El bloque de modelos lineales no lo trata con mucho detalle.

Asin *et al*. [6]. Libro de problemas resueltos orientados a la ingeniería, algunos de ellos con aplicaciones informáticas.

Ayala (*http://www.uv.es/ayala*) [7]. Apuntes de Análisis de datos con R para Ingeniería Informática, que pese a que muchos de los contenidos no se incluyen en los descriptores de esta asignatura, merece la pena leerlo (o al menos disponer de él para futuros problemas con los que se enfrenten los estudiantes), por su completitud, aplicabilidad práctica y su sinceridad,

que lo hace muy ameno. Además, para los alumnos no supone coste alguno al estar disponible en la *web*.

Cuadras [19, 18]. Obra general de problemas y recordatorios de teoría. El primer volumen está dedicado a problemas de probabilidad, mientras que el segundo a problemas de estadística.

Engineering Statics handbook (on-line). Libro dedicado a la Estadística para ingenieros, es bastante amplio y cuenta con la ventaja de estar disponible en la red: http://www.itl.nist.gov/div898/handbook/.

Gonick y Smith [33]. Es un libro que abarca la mayoría de los contenidos de la asignatura, y los explica con ilustraciones simples, muy claras y divertidas, como en un cómic, tal y como indica el título del libro.

Jain [37]. Libro amplio, que abarca también puntos que no entran en el temario, pero que tiene muchas aplicaciones en la informática. Es una referencia completa sobre el análisis del funcionamiento de sistemas informáticos.

Jaisingh [38]. Es un libro de introducción a la estadística, con una presentación muy sencilla y gráfica y con numerosas cuestiones de autoevaluación, de verdadero/falso, elección múltiple y a completar.

Moore [51]. Es la versión castellana de *The Basic Practice of Statistics* [50], donde se realiza una introducción a la estadística, dando gran importancia al trabajo con datos, tal y como recomendaba un comité de la Sociedad Americana de Estadística (ASA) y la Asociación Americana de Matemáticas (MAA) creado para estudiar la enseñanza de la introducción a la Estadística [15].

Navarro *et al.* [44]. Libro de problemas resueltos de probabilidad y estadística para una asignatura introductoria de estadística en diversas titulaciones científico-técnicas.

Ríos [64]. Libro general, clásico y de un cierto nivel.

Rodríguez *et al.* [65]. Libro de ejercicios resueltos con el *Statgraphics*.

Spiegel [73]. Libro general de cierto nivel y con numerosos ejercicios resueltos.

Trivedi [74]. Es una obra de cierto nivel, con aplicaciones informáticas. Cuenta con diversos capítulos dedicados a la fiabilidad y teoría de colas. Sin embargo, el bloque de modelos lineales lo trata muy someramente.

- Para ampliar la primera parte referente a la inferencia estadística, se puede consultar DeGroot [21], también Rohatgi [66] y Vélez y García [76]. Para profundizar en la inferencia bayesiana los siguientes textos pueden ser recomendados: Box [8] y Lindley [42]. Por último, para ampliar la parte de Estadística no Paramétrica podemos consultar Noether [54] y Conover [16].

- Dos excelentes libros que podríamos recomendar para ampliar el apartado de Control de Calidad serían principalmente Montgomery [46] y Prat *et al*. [58].

- Dos libros clásicos sobre Modelos Lineales son Searle [71] y Rao [61]. Más actuales y muy claros son Christensen [14] y Jørgensen [40]. Un libro clásico sobre Anova es Scheffé [70]. Para profundizar en el diseño de experimentos podemos consultar Box *et al*. [9]. Además, dos libros de Montgomery enfocados a la Ingeniería y dedicados a los Diseños de Experimentos y el Análisis de Regresión respectivamente, serían [47] y [48].

- Por último, los siguientes son buenos libros, especializados en estadística usando el programa R o el S-Plus, aunque en la mayoría de los casos los contenidos sobrepasan los objetivos del curso: Selvin [72], Verzani [78] (http://www.math.csi.cuny.edu/Statistics/R/simpleR/index.html), Faraway [28], Dalgaard [20] (http://www.biostat.ku.dk/∼pd/ISwR.html), Venables y Ripley [77] y Everitt [27].

- Fuera ya de lo que son los contenidos de la asignatura, pero que trata el *aprendizaje a partir de los datos*, en una época en la que precisamente, hay un *superávit* de datos, es el magnífico libro de Hastie *et al*. [35], que además cuenta con muchísimas aplicaciones reales a problemas informáticos-estadísticos, desde por ejemplo, el reconocimiento de caracteres al ordenamiento de páginas de Google.

8.3. Material *on-line*

Pese a que esta sección, puede quedarse obsoleta por la rapidez de los cambios en la red, no está de más echar un pequeño vistazo a los innumerables y muy buenos recursos de los que disponemos en internet.

En el departamento de Matemáticas de la Universidad de A Coruña existen numerosos enlaces a webs interesantes sobre docencia en Estadística, con enlace: http://www.udc.es/dep/mate/Dpto_Matematicas/Enlaces/rec_est.htm. En la Universitat Oberta de Catalunya también puede encontrarse buen material y más enlaces interesantes (http://www.uoc.edu/in3/e-math/). La página http://onlinestatbook.com/rvls.hmtl está dedicada a mostrar diversos conceptos mediante simulaciones. También en este sentido, el material suplementario $on-line$ del libro de Moore [51] es muy didáctico. Para complementos, la página (http://www.itl.nist.gov/div898/ handbook/ index.htm) del *Engineering Statistics Handbook* es una buena opción. Es posible encontrar gran cantidad de libros on-line en la página http://digital.library.upenn.edu/webbin/ book/subjectstart?Q, que pueden ser de gran interés por su accesibilidad y su especialización, son varios los libros dedicados a repasar el papel de la Estadística en varias áreas informáticas, como puede ser el *Statistical Software Engineering*. En la página http://www-groups.dcs.st-and.ac.uk/∼history puede encontrarse la biografía de los más ilustres estadísticos y matemáticos. Un *blog* en

castellano con muchos puntos interesantes es http://predictive.wordpress.com/. De hecho, éste y otros enlaces, fueron suministrados por los propios estudiantes, a raíz de una actividad propuesta al inicio del curso, sobre búsqueda de información en la red.

Hay también muchas páginas con iniciativas basadas en las nuevas tecnologías como las webquest (en castellano en http://www.estadisticaparatodos.es hay diverso material interesante), o las wikis, para aprendizaje colaborativo. En este sentido, dirigiéndonos a los docentes, en las distintas revistas sobre Educación estadística, hay muchos artículos interesantes y en muchos casos con los datos disponibles, que sirven para tomar ideas: *Journal of Statistics Education* (http://www.amstat.org/publications/jse/jse index.html), Teaching Statistics (www.rsscse.org.uk/ts/), *Technology Innovations in Statistics Education* (http://repositories.cdlib.org/uclastat/cts/tise/), *Statistics Education Research Journal* (www.stat.auckland.ac.nz/iase/publications.php?show=serj), o *Case Studies in Business, Industry and Government Statistics* (CS-BIGS) (www.bentley.edu/csbigs). Un libro que también recoge gran cantidad de ideas y proyectos para fomentar la participación de los estudiantes es el de Gelman y Nolan [32], aunque ya advierte que para ser efectivas, las clases no deben ser numerosas.

Bibliografía

[1] A. AGRESTI. *An introduction to categorical data.* Wiley, segunda edición, 2007.

[2] A. AGRESTI Y C. FRANKLIN. *Statistics: The Art and Science of Learning from Data.* Prentice Hall, 2006.

[3] A. O. ALLEN. *Probability, Statistics and Queueing Theory with Computer Science Applications.* Academic Press, 1990.

[4] F. J. ALONSO, P. A. GARCÍA Y J. E. OLLERO. *Estadística para ingenieros (teoría y problemas).* Colegio de Ingenieros de Caminos, Canales y Puertos, 1996.

[5] R. ARDANUY Y Q. MARTÍN. *Estadística para ingenieros.* Editorial Hespérides, 1993.

[6] J. ASÍN, F. G. BADÍA, M. D. BERRADE, C. A. CAMPOS, C. GALÉ Y P. JODRÁ. *Probabilidad y estadística en ingeniería: ejercicios resueltos.* Prensas universitarias de Zaragoza, 2002.

[7] G. AYALA. Apuntes de análisis de datos con R para ingeniería informática. http://www.uv.es/ayala.

[8] G. E. P. BOX Y N. DRAPER. *Bayesian Inference in Statistical analysis.* Wiley, 1992.

[9] G. E. P. BOX, W. G. HUNTER Y J. S. HUNTER. *Estadística para investigadores.* Editorial Reverté, 1993.

[10] G. C. CANAVOS. *Probabilidad y Estadística. Aplicaciones y métodos.* McGraw-Hill, 1993.

[11] R. CAO, M. FRANCISCO, S. NAYA, M. A. PRESEDO, M. VÁZQUEZ, J. A. VILAR, Y J. M. VILAR. *Introducción a la estadística y sus aplicaciones.* Pirámide, 2001.

[12] M. A. CASTRO Y Y. VILLACAMPA. *Estadística aplicada a la ingeniería civil.* Club Universitario, 2000.

[13] C. CHATFILED. *Statistics for technology.* Chapman and Hall, 1983.

[14] R. CHRISTENSEN. *Plane answer to complex questions. The theory of linear models.* Springer, 1987.

[15] G. COBB. Teaching statistics. In *Heeding the Call for Change: Suggestions for Curricular Action*, pages 3–43. MAA, 22, Notes Mathematical Association of America, 1992.

[16] W. J. CONOVER. *Introduction to statistics: a nonparametric approach.* Wiley, 1976.

[17] J. L. CORONADO, A. CORRAL, P. LÓPEZ, R. MIÑANO, B. RUIZ Y J. VILLÉN. *Estadística aplicada con STATGRAPHICS.* Ra-ma, 1994.

[18] C.M. CUADRAS. *Problemas de probabilidades y estadística Vol. 2, Inferencia estadística.* PPU, 1991.

[19] C. M. CUADRAS. *Problemas de probabilidades y estadística Vol. 1, Probabilidades.* EUB, 1999.

[20] P. DALGAARD. *Introductory Statistics with R.* Springer, 2002.

[21] M. H. DEGROOT. *Probabilidad y Estadística.* Addison-Wesley Iberoamericana, 1988.

[22] J. L. DEVORE. *Probabilidad y estadística para ingeniería y ciencias.* International Thomson, cuarta edición, 1998.

[23] J. DOMINGO. *Estadística tècnica. Una introducció contructivista.* Universitat Rovira i Virgili, segunda edición, 1997.

[24] E. R. DOUGHERTY. *Probability and statistics for the engineering, computing and physical sciences.* Prentice Hall Internatinal Editions, 1990.

[25] J. M. EGUSQUIZA. *Apuntes de métodos estadísticos de la ingeniería.* Geneve, 1998.

[26] I. EPIFANIO Y A. RÓDENAS. *Material docente para prácticas con R de la asignatura IG23 Ampliación de Estadística.* Publicacions de la Universitat Jaume I, 2006.

[27] B. S. EVERITT. *A handbook of statistical analyses using S-PLUS.* Chapman & Hall, 1994.

[28] J. FARARWAY. *Linear models with R.* CRC Press, 2002.

[29] F. FERNÁNDEZ, M. A. LÓPEZ, M. MÚÑOZ, A. M. RODRÍGUEZ, A. SÁNCHEZ, Y C. VALERO. *Estadística asistida por ordenador STATGRAPHICS PLUS 4.1.* Universidad de Cádiz, 2000.

[30] A. GÁMEZ Y L. M. MARÍN. *Estadística para ingenieros técnicos.* Universidad de Cádiz, 2000.

[31] A. GARCÍA, V. HERNÁNDEZ, H. NAVARRO, E. RAMOS, R. VÉLEZ Y I. YÁÑEZ. *Estadística I. Ingeniería Técnica en Informática de Sistemas.* UNED, 1994.

[32] A. GELMAN Y D. NOLAN. *Teaching Statistics: a bag of tricks.* Oxford University Press, 2002.

[33] L. GONICK Y W. SMITH. *La Estadística en cómic.* Editorial Zendrera Zariquiey, 2002.

[34] P. GREGORI E I. EPIFANIO. *Estadística bàsica per a la titulació d'Enginyeria Tècnica en Informàtica de Gestió: teoria i pràctiques amb el programa R.* Publicacions de la Universitat Jaume I, en revisión, 2008.

[35] T. HASTIE, R. TIBSHIRANI Y J. FRIEDMAN. *The Elements of Statistical Learning. Data mining, inference and prediction.* Springer-Verlag, segunda edición, 2009.

[36] V. HERNÁNDEZ, E. RAMOS E I. YÁÑEZ. *Estadística I. Ingeniería Técnica en Informática de Gestión.* UNED, 1994.

[37] R. JAIN. *The art of computer systems performance.* Wiley, 1991.

[38] L. R. JAISINGH. *Statistics for the utterly confused.* McGraw-Hill, 2000.

[39] R. A. JOHNSON. *Probabilidad y estadística para ingenieros de Miller y Freund.* Prentice Hall Hispanoamericana, quinta edición, 1997.

[40] B. JØRGENSEN. *The theroy of linear models.* Chapman and Hall, 1993.

[41] F. LEISCH. Dynamic generation of statistical reports using literate data analysis. In *Compstat 2002 - Proceedings in Computational Statistics.* Physika Verlag, 2002. http://www.ci.tuwien.ac.at/∼leisch/Sweave.

[42] D. W. LINDLEY. *Introduction to Probability and Statistics from a Bayesian Viewpoint. (2 vol.).* Cambridge University Press, 1969.

[43] A. LLORIA, I. EPIFANIO Y A. BELTRÁN. *Material docente para el autoaprendizaje y la autoevaluación de la asignatura IS12 Estadística.* Publicacions de la Universitat Jaume I, 2003.

[44] J. NAVARRO, M. FRANCO Y A. GUILLAMÓN. *Probabilidad y estadística. Problemas.* Diego Marín, 1999.

[45] W. MENDENHALL Y T. SINCICH. *Probabilidad y estadística para ingeniería y ciencias.* Prentice Hall, cuarta edición, 1997.

[46] D. C. MONTGOMERY. *Introduction to Statistical Quality Control.* John Wiley and Sons, 1985.

[47] D. C. MONTGOMERY. *Diseño y análisis de experimentos.* Grupo Editorial Iberoamérica, tercera edición, 1991.

[48] D. C. MONTGOMERY Y E. A. PECK. *Introduction to Linear Regression Analysis*. John Wiley and Sons, 1992.

[49] D. C. MONTGOMERY Y G. C. RUNGER. *Probabilidad y estadística aplicadas a la ingeniería*. Limusa Wiley, segunda edición, 2002.

[50] D. S. MOORE. *The Basic practice of statistics*. Freeman, 1995.

[51] D. S. MOORE. *Estadística aplicada básica*. Antoni Bosch Editor, traducción y adaptación de Jordi Comas, 1998.

[52] M. A. MORALES. Generación automática de reportes con R y LATEX. Technical report, http://cran.r-project.org/doc/contrib/Rivera-Tutorial_Sweave.pdf, 2006.

[53] M. J. NICOLÁS. *Estadística aplicada con R*. Nausícaä, 2003.

[54] G. E. NOETHER. *Practical Nonparametric Statistics*. Houghton Mifflin Co., 1980.

[55] D. PEÑA. *Estadística. Modelos y métodos*, volumen 2. Modelos lineales y series temporales. Alianza Editorial, segunda edición, 1989.

[56] D. PEÑA. *Estadística. Modelos y métodos*, volumen 1. Fundamentos. Alianza Editorial, segunda edición, 1991.

[57] C. PÉREZ. *Estadística práctica con StatGraphics*. Prentice Hall, 2001.

[58] A. PRAT, X. TORT-MARTORELL, P. GRIMA Y L. POZUETA. *Métodos estadísticos. Control y mejora de la calidad*. Universitat Politècnica de Catalunya, segunda edición, 1995.

[59] W. H. PRESS, B. P. FLANNERY, S. A. TEULOSKY Y W. T. VETTERLING. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, 1992.

[60] R DEVELOPMENT CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2009. ISBN 3-900051-07-0.

[61] C. R. RAO. *Linear Statistical Inference and its Applications*. Wiley, second edición, 1973.

[62] A. RAS. *Estadística aplicada per a enginyeria*. Edicions UPC. Universitat Politécnica de Catalunya, 1994.

[63] A. RAS, G. OLIVAR Y D. MARTIN. *Estadística. Pràctiques*. Universitat Politècnica de Catalunya, 1993.

[64] S. RÍOS. *Métodos estadísticos*. Ediciones del Castillo, 1977.

[65] R. RODRÍGUEZ, A. INFANTE, J. VALDIVIESO Y M. FERNÁNDEZ. *Estadística práctica con StatGraphics*. Universidad de Cádiz, 1997.

[66] V. K. ROHATGI. *Statistical Inference*. Wiley, 1984.

[67] R. ROMERO Y L. R. ZÚNICA. *Métodos estadísticos en ingeniería*. Universidad Politécnica de Valencia, 2005.

[68] S. M. ROSS. *A course in simulation*. Prentice Hall, 1990.

[69] R. L. SCHEAFFER Y J. T. McCLAVE. *Probabilidad y estadística para ingeniería*. Grupo Editorial Iberoamérica, 1993.

[70] H. SCHEFFÉ. *The Analysis of Variance*. John Wiley and Sons, 1959.

[71] S. R. SEARLE. *Linear Models*. Wiley, 1971.

[72] S. SELVIN. *Modern applied biostatistical methods using S-PLUS*. Oxford University Press, 1998.

[73] M. R. SPIEGEL. *Estadística*. McGraw-Hill/ Interamericana de España S.A., 1991.

[74] K. S. TRIVEDI. *Probability and Statistics with Reliability, Queueing and Computer Science Applications*. Prentice-Hall, 1982.

[75] M. D. UGARTE Y A. F. MILITINO. *Estadística aplicada con S-PLUS*. Universidad Pública de Navarra, 2002.

[76] J. R. VÉLEZ Y A. GARCÍA. *Principios de Inferencia Estadística*. UNED, 1993.

[77] W. N. VENABLES Y B. D. RIPLEY. *Modern applied statistics with S-PLUS*. Springer, 2002.

[78] J. VERZANI. *Using R for introductory statistics*. Chapman & Hall, 2005.

[79] J. M. VILAR. *Modelos estadísticos aplicados*. Universidade da Coruña, 2003.

[80] R. E. WALPOLE, R. H. MYERS Y S. L. MYERS. *Probabilidad y estadística para ingenieros*. Prentice Hall, sexta edición, 1998.